Deciphering context-dependent epigenetic program by network-based prediction of clustered open regulatory elements from single-cell chromatin accessibility

Running title: *Park et al.*; *Network-based prediction of clustered open regulatory elements at the single-cell level*

Seonjun Park, BSc[1]; Sungkook Ma, BSc[1], Wonhyo Lee, BSc[1], Sung Ho Park, PhD[1*]

[1] Department of Biological Sciences, Ulsan National Institute of Science & Technology (UNIST), Ulsan, Korea.

Correspondence should be addressed to: [*] Department of Biological Sciences, Ulsan National Institute of Science & Technology (UNIST), Ulsan, Korea.

parksu@unist.ac.kr

Abstract

Large cis-regulatory domains, such as super-enhancers, are pivotal in orchestrating robust and cell-state-specific transcriptional programs that define cellular identity. However, current single-cell methods do not effectively identify these higher-order structures, obscuring the coordinated, domain-level epigenetic regulation essential for complex biological processes. Identifying such domain-scale representation at the single-cell level is critical for understanding the regulatory logic underlying development and disease. Here, we introduce enCORE, a computational framework that leverages enhancer-enhancer interaction networks to determine Clustered Open Regulatory Elements (COREs) from single-cell ATAC-sequencing. Our approach faithfully recapitulated established hematopoietic hierarchies and resolved lineage-specific regulatory programs by recovering canonical lineage-defining regulators, frequent chromatin interactions, and enrichment of fine-mapped autoimmune disease-associated genome-wide association study (GWAS) variants. In colorectal cancer, enCORE successfully captured tumor-associated H3K27ac programs and prioritized cancer-relevant regulators, pointing to *USP7* as a potential therapeutic candidate supported by *in silico* perturbation. Our framework provides a fruitful approach for deciphering context-dependent epigenetic reprogramming.

Main

Gene expression programs are governed not only by individual regulatory elements, but also by complex interactions within large cis-regulatory domains. These domains, spanning tens to hundreds of kilobases, function as regulatory hubs integrating multiple inputs to ensure robust and precise transcriptional output. Their architectures are diverse, including activating domains like super-enhancers (SEs) [1, 2], stretch enhancers [3], broad H3K4me3 domains [4], and repressive domains such as super-silencers [5, 6], Lamin-associated domains (LAD) [7]. These domains form the functional backbone of the non-coding genome, enabling context-dependent transcriptional regulation, cell identity specification, and disease-associated phenotypes.

Among these large cis-regulatory domains, highly interactive enhancer clusters, recognized as SEs or highly inter-connected enhancers [8], are well-established mediators of stable transcriptional activation that orchestrate robust expression of lineage-determining and disease-associated genes. These enhancer clusters are marked by three key characteristics: (1) Coordinated chromatin accessibility associated with broad H3K27ac domains [9-11], (2) Dense transcription factor (TF) occupancy, facilitating cooperative assembly of regulatory complexes, and (3) Frequent enhancer-enhancer chromatin interactions [8, 12], sustaining spatial proximity and potent transcriptional output. At the genetic level, genome-wide association studies (GWAS) variants are significantly enriched within these enhancer clusters [2], providing a mechanistic context for noncoding disease risk loci.

A major objective in studying these enhancer clusters is their profound cellular specificity. The activity and conformation of such clusters are highly dynamic among cell types, developmental stages, and disease states. Traditionally, these domains have been studied using bulk-level assays, such as ChIP-seq, yet these approaches average signals across heterogeneous cells and mask critical cell-specific regulation. Therefore, resolving these domains at the single-cell level is essential to stratify heterogeneous populations, allowing us to decode the spatially coordinated chromatin features and dense regulatory complexes specific to distinct cell states that contribute to complex tissue function and disease.

Single-cell ATAC sequencing (scATAC-seq) [13] has emerged as a powerful tool to investigate these epigenetic landscapes while resolving cellular heterogeneity. scATAC-seq provides a comprehensive

map of cis-regulatory elements for many cell types simultaneously [14-16]. For instance, scATAC-seq-based chromatin accessibility maps can delineate cell type-specificity and help interpret noncoding variants associated with complex human traits [15]. In this way, scATAC-seq can be utilized to identify the master regulatory elements that may be crucial in various biological contexts, such as lineage-specific or disease-specific scenarios, thereby broadening our understanding of genome-wide epigenetic mechanisms. Additionally, since scATAC-seq is experimentally tractable and reproducible at scale, it has been widely adopted in single-cell studies, providing an unprecedented opportunity to map highly interactive enhancer clusters onto specific cell states.

However, despite the clear biological importance of highly interactive enhancer clusters, no computational framework currently exists to predict such domain-level clusters based solely on scATAC-seq data. Instead, current standard workflows focus on discrete peaks, meaning that most scATAC-seq analyses remain element-centric. This approach makes domain-scale cooperativity difficult to observe directly and fragments enhancer clusters into isolated features. Thus, there is a need for a methodological approach to infer highly interactive enhancer clusters from scATAC-seq data.

In the present study, we develop enCORE, a framework to predict clustered open regulatory elements (CORE) from scATAC-seq data by integrating three major characteristics of highly interactive enhancer clusters, chromatin accessibility, chromatin interactions inferred by co-accessibility, and TF motif information. We discovered that enCORE recapitulates cell type-specific transcriptional regulation in peripheral blood (PB). In addition, it has been shown that both enhancer-enhancer chromatin interactions and immune-related disease-associated GWAS variants were enriched within CORE. In the bone marrow (BM), we demonstrated that enCORE effectively captures epigenetic states associated with hematopoietic lineage commitment. enCORE further supported a lineage-matched maturation continuity from BM to PB and suggested that macrophage-biased H3K27ac was partially anticipated by signatures present in early monocyte states. We showed that enCORE elucidates colorectal cancer (CRC)-related epigenetic reprogramming, thus providing new insight into underlying disease mechanisms and identifying potential therapeutic candidates.

Results

Overview of enCORE algorithm

We developed enCORE, a computational framework for identifying highly interactive enhancer clusters from scATAC-seq data (Fig. 1a). First, we defined the enhancer clusters by several data processing steps: (1) Extracting the enhancer candidates and constructing the initial enhancer module, (2) Merging two nearest, disconnected enhancer module with similar inferred gene-regulatory relationships (Link loss correction), (3) Estimating a one-dimensional genomic distance threshold for valid interactions to partition merged enhancer module (Distance threshold Optimization), (4) Additional rescue steps for compensating the sparsity (Sparsity correction) (Methods). The framework then constructs a pseudo-directed enhancer-enhancer interaction network, assigning node and edge weights using accessibility, TF weight, co-accessibility (correlation), and genomic distance (Methods; Fig. 1a,b). Finally, the personalized PageRank algorithm was adopted to rank enhancer clusters. There are two options in enCORE: potential and active. The potential option corresponds to the original enCORE implementation and may include clusters with characteristics of poised/primed states. In contrast, the active option applies promoter accessibility-based filtering to exclude clusters likely reflecting poised/primed states, thereby emphasizing regulatory programs engaged in the current cell state (Methods).

To validate the enCORE framework, we utilized three scATAC-seq datasets: human peripheral blood mononuclear cells (PBMC), bone marrow mononuclear cells (BMMC) [17], and colorectal cancer (CRC) [18] (Methods; Fig. 1c). We extracted CORE from these three scATAC-seq datasets (Supplementary Table 1-3). We also collected publicly available fine-mapped GWAS variants from CausalDB [19], curated GWAS variants from Hnisz *et al.* [2], SEs (in CD4+ T, CD8+ T, NK, B, and CD14+ Monocytes) from H3K27ac bulk ChIP-seq data in SEdb [20], Hi-C for CD19+ B cells and CD14+ Monocytes [21], HiChIP for H3K27ac from HiChIPdb [22], and metastasis-related genes from metsDB [23] (Methods).

Cell type-specific transcriptional regulation of CORE-associated genes in PBMC

We first applied enCORE to the 10k PBMC scATAC-seq dataset, identifying COREs in seven immune cell types: CD4+ Memory T cells, CD8+ Effector T cells, B cells, NK cells, CD14+ and CD16+ Monocytes, and classical Dendritic cells (Supplementary Table 1). We hypothesized that CORE reflects cell identity

and cell-type specificity, similar to SE [2]. To evaluate whether CORE shows SE-like cell-type specificity relative to typical enhancers within the same active-enhancer background, we contrasted CORE with non-CORE active enhancers (non-CORE), defined as the merged region of SEs and typical enhancers (TEs) that do not overlap with CORE and compared the two groups. Genes associated with CORE were highly cell-type specific relative to non-CORE-associated genes (Fig. 2a). This trend was consistently observed under both the potential and active options, analogous to the SE versus TE comparison (Extended Data Fig. 1a,b). Enhancer candidates inferred by enCORE showed substantial overlap with H3K4me1 peaks in each cell type, supporting the validity of our algorithm in capturing enhancer activity (Fig. 2b). Among CORE-associated genes, we detected cell type-specific master regulators, including *GATA3* in NK/T cells, *EBF1* in B cells, and *MAFB* in myeloid cells, each showing association with CORE in the corresponding cell types (Fig. 2c).

To evaluate whether CORE-associated genes represent cell-type-specific transcriptional patterns, we utilized publicly available PBMC RNA-seq datasets (Methods). Similar to the trend observed between SE and TE, CORE-associated genes displayed more cell type-specific transcriptional activation than non-CORE-associated genes (Fig. 2d, Extended Data Fig. 1c). Expression patterns of CORE-associated hematopoietic master regulators (*GATA3*, *EBF1*, *MAFB*) also showed cell type specificity (Fig. 2e). TF motif enrichment analysis further confirmed that CORE exhibited stronger enrichment for cell type-specific TF motifs compared with non-CORE regions (Fig. 2f). These results support that enCORE can recapitulate cell type-specific transcriptional regulation in PBMC.

enCORE is designed to capture domain-level cooperation by integrating TF motif and network information, while SEs are typically defined based on coactivator-associated ChIP-seq signal intensity (e.g., H3K27ac or Mediator). Since the prioritization criteria are different, we hypothesized that CORE would be distinguishable from SE while significantly overlapping with it. ROC curve analysis revealed that CORE substantially overlaps with SE but represents different profiles (Extended Data Fig. 1d,e and Supplementary Table 4). GO enrichment analysis was performed in CD14+ and CD16+ Monocytes showed that while both CORE and SE were enriched for cell type-specific terms relative to TE, CORE retained a subset of genes not captured by SE (Supplementary Fig. 1a,b). For example, in CD14+ Monocytes, the *KLF4* gene (macrophage polarization [24]) and the *RBPJ* gene (monocyte-to-macrophage differentiation [25]) were targeted only by CORE (Supplementary Fig. 2a,b). In contrast,

the *LYZ* gene (canonical marker of monocytes [26]) was exclusively identified by SE (Supplementary Fig. 2c). These observations indicate that CORE and SE emphasize shared, but distinct layers of regulatory features within the same cell population.

We next evaluated whether SE defined from H3K27ac single-cell CUT&Tag-seq (scCUT&Tag-seq) [27] data detect highly interactive enhancer clusters to an extent comparable to CORE from scATAC-seq. We identified SEs from the PBMC H3K27ac scCUT&Tag-seq dataset [28] using the ROSE algorithm and compared them with both CORE and SE (Extended Data Fig. 2a,b, Supplementary Table 5). The assumption that these enhancer clusters can be detected equally well using scCUT&Tag-seq targeting H3K27ac may undermine the distinct advantages and contributions of the enCORE framework. Given the sparsity of scCUT&Tag-seq data [27], we hypothesized that scCUT&Tag-derived SE would perform suboptimal compared with CORE or SE. scCUT&Tag-derived SE showed cell-type-specific annotation patterns similar to CORE and SE (Extended Data Fig. 2c). However, scCUT&Tag-derived SE underperformed in identifying hematopoietic master regulators compared with CORE and SE (Extended Data Fig. 2d). Although scCUT&Tag-derived SE showed a trend toward higher mRNA-level cell type specificity than TE from the same dataset, the difference was not statistically significant (Extended Data Fig. 2e). Collectively, these observations underscore the practical utility of scATAC-based enCORE, which more robustly recovers highly interactive enhancer clusters than H3K27ac scCUT&Tag-derived SE under technically sparse conditions.

Enhancer constituents within CORE frequently interact with other enhancers

We assessed whether CORE faithfully captures the frequent chromatin interactions underlying highly interactive enhancer clusters, a functional organization known to support robust, cell state-specific transcription. We observed a significant enrichment of chromatin interactions within B cell-derived CORE constituents compared with non-CORE regions in both the H3K27ac HiChIP (GM12878) and Hi-C (CD19+ B cells) datasets (Fig. 3a,b). Similar enrichment was validated in CD4+ T cells and CD14+ Monocytes (Extended Data Fig. 3a,b). These observations suggest that CORE preferentially identifies chromatin interaction-enriched enhancer architectures across diverse immune cell types.

Next, given that differences in enhancer activity could confound comparisons of context-dependent interaction strength [29], we first characterized the epigenetic states of enhancer constituents within CORE and SE. Using the ChIP-seq datasets for H3K27ac and H3K27me3, we categorized H3K4me1-marked enhancer peaks into three state categories: active, poised, and primed enhancers (Methods). Compared with SE, CORE had a higher proportion of poised enhancers (Fig. 3c). Although using the active option in a CORE calling may result in partial loss of active enhancer information, the proportion of active enhancers relative to poised enhancers increased (Supplementary Fig. 3a,b). These results suggest that the CORE tends to capture a greater number of transcriptionally poised epigenetic states compared with SE, hinting at a higher sensitivity to future-responsive regulatory potential. Moreover, we demonstrated that the active option effectively removes these poised enhancers during CORE calling.

We then subsequently compared enhancer-enhancer chromatin interactions between CORE and SE stratified by epigenetic state categories. In CD19+ B cells, CORE showed consistently stronger enhancer-enhancer chromatin interactions across all three enhancer categories compared with SE (Fig. 3d). Notably, only CORE was detected near the *PRDM1* gene, a key regulator of B cell-to-Plasma differentiation [30, 31], exhibiting strong internal interactions (Fig. 3e). In CD14+ Monocytes, CORE showed a mild but insignificant increase in the degree of interaction of poised enhancers, and in the rest of the categories, it showed similar levels to SE (Extended Data Fig. 3c). Taken together, these results demonstrate that enCORE captures enhancer-associated chromatin interaction architectures comparably to SE in the single-cell setting.

Enrichment of immune-related disease-associated GWAS variants within CORE

We investigated whether fine-mapped GWAS variants are enriched within CORE. To measure the enrichment of GWAS variants, we defined an enrichment score as the fold change of GWAS variants within the enhancer group relative to genomic background (Methods). Both CORE and SE showed a pronounced enrichment of fine-mapped GWAS variants associated with immune-related diseases compared with those linked to non-immune traits (Fig. 4a, Extended Data Fig. 4a, Supplementary Fig. 4a). We also observed disease-specific enrichment patterns consistent with known disease-critical cell

types [32-35]. The fine-mapped GWAS variants for Systemic Lupus Erythematosus (SLE) and Chronic Lymphocytic Leukemia (CLL) were enriched within B cell-derived CORE, whereas the variants for Asthma and Type 1 Diabetes (T1D) were mainly found within T cell-derived CORE (Fig. 4b,c, Extended Data Fig. 4b-d, Supplementary Fig. 4b,c). In Multiple Sclerosis (MS), variants were more enriched in CD8+ T cells and B cells than in CD4+ T cells (Extended Fig. 4e, Supplementary Fig. 4d). This is consistent with reports on the clinical failure of CD4+ T cell-targeted therapy in MS, and the important role of CD8+ T cells and B cells in MS pathogenesis [36, 37]. This trend was validated across multiple GWAS cohorts (Extended Data Fig. 4, Supplementary Fig. 4). Additionally, motivated by the lipid-response signature observed in monocytes-derived CORE (Supplementary Fig. 1a,b), we next examined whether coronary artery disease (CAD) risk loci show preferential enrichment in monocytes. In three GWAS cohorts, CAD risk loci were most strongly enriched in CD14+ Monocytes (Extended Data Fig. 5a-d). A representative fine-mapped CAD risk loci near the *LIPA* was recovered only by CORE (Extended Data Fig. 5e). Given the role of *LIPA* in lipid metabolism [38] and the reported association between CAD risk alleles near *LIPA* and its mRNA levels in monocytes [39], this finding implicates CORE-associated regulation of *LIPA* as a candidate mechanism contributing to CAD risk.

Our previous observation that CLL risk loci are enriched within B cell-derived CORE underscores B cell-specific nature of CLL initiation (Fig. 4c). We hypothesized that CORE would capture CLL post-initiation transcriptional changes more prominently in line with a putative advantage in reflecting future-responsive transcriptional potential compared with SE. As expected, more differentially expressed genes were detected in CORE than SE (Fig. 4d). This trend was seen in both CORE options, and the potential option indicated a higher number of up-regulated genes. Among up-regulated genes, *RXRA* has been exclusively identified by CORE (Fig. 4e). Intriguingly, *RXRA* gene expression has been reported to be significantly increased in both human and mouse CLL cells compared with normal B cells [40], suggesting the potential role of epigenetic regulation of *RXRA* in CLL pathogenesis. To sum up, enCORE provides a cell-type-resolved framework to interpret fine-mapped GWAS variants by linking noncoding disease risk loci to domain-level regulation and candidate target genes.

enCORE reconstructs hematopoietic lineage trajectories and lineage-defining

regulators in BMMC

To evaluate whether CORE profiles encode hematopoietic lineage commitment, we applied enCORE to a BMMC scATAC-seq dataset (Supplementary Table 2). We quantified cross-cell-type similarity of CORE profiles using cosine similarity to reconstruct epigenetic proximity along the hematopoietic hierarchy (Supplementary Table 6a). The resulting patterns summarized the established hematopoietic lineage trajectory (Fig. 5a, Extended Data Fig. 6a,b). We observed an association among Granulocyte-Myeloid-Progenitors (GMP), Common Lymphoid-Progenitors (CLP), and plasmacytoid Dendritic Cells (pDC), supporting reports that pDC can originate from both lymphoid and myeloid lineages [41]. This finding appears to be attributable to the nature of the potential option that is expected to reflect more poised regulatory states than the active option.

Building on the observation that CORE profiles track hematopoietic lineage trajectory, we next investigated whether CORE-associated genes recover known lineage-defining regulators. We identified *EBF1* (B cell), *GATA3* (NK/T cell), *MAFB* (Monocyte/Myeloid), and *GATA1* (Megakaryocyte/Erythroid) in their respective lineage (Fig. 5b). *TCF4* was exclusively observed in pDC, except for the B cell lineage. In Hematopoietic Stem Cells (HSC), *MECOM* (*EVI1*) and *PRDM16* were detected. Both genes are involved in self-renewal [42] and quiescence [43], respectively. Notably, *MECOM* was also observed in both Megakaryocyte-Erythroid-Progenitors (MEP) and CLP. This result may be related to reports of *MECOM*-associated mutations that cause amegakaryocytic thrombocytopenia and B cell deficiency [44]. These observations underscore the utility of enCORE in linking lineage-resolved regulatory programs to hematopoietic phenotypes.

Despite clear lymphoid/myeloid lineage discrimination, CORE profiles between CLP and GMP were similar under the potential option. Considering that GMP in our dataset is *GATA1*-negative, our findings support a previously proposed model [45] in which *GATA1*-negative GMP is derived from Lymphoid-Myeloid-Primed-Progenitors (LMPP) and shares a trajectory with the lymphoid lineage. To examine whether GMP/granulocytes originated primarily from LMPP, we applied MultiVelo [46] to CD34+ hematopoietic cells (Fig. 5c). At least in this dataset, it was inferred that GMP/granulocytes derived mainly from LMPP. Based on these results, we have outlined a model of bone marrow hematopoiesis (Fig. 5d). Collectively, enCORE summarizes hematopoietic lineage structure and lineage-defining regulators at the epigenetic level, providing a practical framework for studying developmental

progression across hematopoiesis.

We additionally asked whether CORE profiles preserve maturation continuity across compartments by comparing matched lineages between bone marrow (BM) and peripheral blood (PB). As expected, PB CORE profiles most closely matched their late-stage counterparts within the BM hierarchy, and the within-lineage nearest-neighbor ordering was preserved (Extended Data Fig. 7a and Supplementary Table 6b). For example, CORE profile derived from PB B cells showed the highest cosine similarity to BM immature B cells, with progressively lower similarity to Pre-B, Pro-B, and CLP states. In myeloid/monocyte lineage, CORE from PB CD14+ Monocytes was most similar to BM CD14+ Monocytes rather than to GMP, and the nearest-neighbor ordering from BM CD14+ Monocytes to PB populations, CD14+ Monocytes followed by CD16+ Monocytes, cDC was preserved. In summary,

The observed preservation of PB epigenetic profiles within the BM hierarchy suggests anticipatory imprinting of downstream fates in hematopoietic cells. We therefore hypothesized that CORE in BM CD14+ monocytes encodes early epigenetic traces of macrophage, beyond PB CD14+ Monocytes. We evaluated the enrichment of overlap between macrophage-monocyte H3K27ac differential peaks and BM-/PB-derived CORE/SE in CD14+ Monocytes (Methods; Extended Data Fig. 7b and Supplementary Table 7). Both CORE and SE preferentially coincide with monocyte-elevated peaks. In contrast, for macrophage-elevated peaks, SE showed negligible enrichment, whereas CORE exhibited statistically significant enrichment. This finding suggests that, at least in monocytes, CORE contains modest but significant macrophage-associated epigenetic signatures, with this tendency apparent beginning in BM.

Genome browser tracks highlighted representative loci where macrophage-elevated peaks overlapped BM-/PB-derived CORE but not SE, including *DCSTAMP* (involved in multinucleation during differentiation toward macrophage-derived giant cells/osteoclasts [47] & in macrophage migration and cytokine production [48]) and *MEF2A* (required for IFN I induction in response to multiple innate immune stimuli [49]) (Extended Data Fig. 7c-e and Supplementary Table 7). Additionally, although peaks near *BHLHE41* fell within the +-2 kb TSS exclusion and were therefore omitted from the enrichment analysis (Methods), they were observed exclusively within CORE. Notably, *BHLHE41* is a key regulator in monocytes-to-alveolar macrophage differentiation [50] and is suggested as an important player in macrophage-specific active epigenetic state marked by H3K27ac [51]. In summary, enCORE partially retains lineage-biased epigenetic signals that are attenuated in SE-based summaries, early fate-

skewed signals become more detectable. This may guide subsequent analyses toward regulatory loci associated with later differentiation programs.

Deciphering CRC-dependent epigenetic reprogramming and potential therapeutic candidates

Lastly, we evaluated the ability of enCORE to detect disease-dependent epigenetic programs in a colorectal cancer (CRC) dataset [18]. We extracted COREs from CRC epithelial cells (Supplementary Fig. 5 and Supplementary Table 3), and SEs from two published H3K27ac ChIP-seq datasets [52, 53] (Supplementary Table 8-10). To compare the performance of CRC-dependent epigenetics program identification, we performed NMF/K-means clustering using the H3K27ac profiles of CRC and normal samples (Methods). CORE distinguished CRC and normal state, whereas SE explained only a small portion of CRC samples (Fig. 6a, Supplementary Fig. 6a,b, Supplementary Fig. 7a,b,g). CORE showed higher log2 fold changes than SE or TE (Fig. 6b, Supplementary Fig. 7h). This trend was further cross-validated by XGBoost classification (Supplementary Fig. 6c-f, Supplementary Fig. 7c-f). We performed *de novo* TF motif enrichment analysis on both CORE and non-CORE constituents. Notably, we discovered the exclusive enrichment of the *FOXM1* TF motif within CORE constituents (Fig. 6c). *FOXM1* is one of the critical regulators in CRC pathogenesis [54, 55]. This suggests that CORE may play a central role in CRC by interacting with cancer master regulators such as *FOXM1* through TF binding.

Next, we investigated whether well-known CRC-related master TF candidates were among CORE- or SE-associated genes (Supplementary Table 11). These candidates were enriched in both CORE- and SE-associated genes (Supplementary Fig. 8). TCGA survival curves were then used to assess whether novel prognosis-related genes could be identified through enCORE. CORE-associated genes, both *EGFR* and *MACC1*, showed significant association with CRC survival (Fig. 6d). *EGFR* is a well-known therapeutic target in CRC [56], and *MACC1* is closely related to CRC metastasis [57, 58]. Other genes, such as *ISM1*, *PRDM15*, and *HOXC10*, which have recently been proposed as novel prognosis, radioresistance, or metastasis-related genes [59-61], have been identified exclusively by CORE (Supplementary Fig. 9). *BRD2* is another gene uniquely identified by CORE, and interestingly, *BRD2* is

listed in the metsDB as one of the candidate genes involved in CRC metastasis (Supplementary Table 12). Considering that *BRD2* functions as an epigenetic reader of histone acetylation [62], it suggests that CORE-mediated regulation of *BRD2* may have a role in metastasis-related epigenetic states of CRC.

We next asked whether CORE is systematically enriched for metastasis-related regulatory programs. Since alterations in three-dimensional (3D) chromatin architectures can potentiate invasion-associated transcriptional programs [63], we examined the enrichment of metastasis-related genes around CORE and its association with enhancer-enhancer interaction network topology. Both metastasis-related TFs and their target genes were more abundant in CORE-associated genes than SE-associated genes (Extended Data Fig. 9a and Supplementary Table 12). We then evaluated node-level properties in the co-accessibility network (Methods). While SE had higher degrees, CORE showed higher clustering coefficients (Extended Data Fig. 9b). This pattern was also observed in an independent PBMC Hi-C interaction network, where CORE showed comparable or slightly lower degrees, but consistently higher clustering coefficients relative to SE (Extended Data Fig. 9c). Peaks with high clustering coefficients showed a lower observed/expected ratio for metastasis-related TFs, but a higher ratio of metastasis-related target genes relative to high-degree peaks (Extended Data Fig. 9d). Notably, despite the reduced relative enrichment for TFs, peaks with high clustering coefficients contained a higher absolute number of both metastasis-related TFs and their target genes, consistent with the patterns observed for CORE versus SE. These results suggest that metastasis-related signatures can also be identified in enCORE-derived locally clustered enhancers, in addition to traditional hub enhancers characterized by a high degree.

While the previous analyses focused on master TF candidates and prognosis/metastasis-related genes, we next explored the therapeutic potential of CORE-associated genes. To prioritize targetable candidates, we focused on CORE-associated genes with reported functional evidence and therapeutic relevance in CRC. In this context, *USP7* was identified exclusively by CORE (Fig. 6e, Extended Data Fig. 8a). Previous studies have suggested *USP7* as a promising therapeutic target [64-66]. Intriguingly, *USP7* knockdown in CRC organoids promoted the reversion of them to a more normal-like state [66]. To further assess whether the *USP7*-proximal CORE bears the putative regulatory influence on *USP7* gene expression, we performed a computational perturbation experiment using AlphaGenome [67] with

the Left colon ontology (Methods; Extended Data Fig. 8b,c and Supplementary Table 13). *In silico* knockout of the H3K27ac enhancer peak corresponding to this CORE predicted a reduction in *USP7* transcript abundance. This *in silico* perturbation result supports the CORE-mediated regulatory potential in modulating the *USP7* gene expression. Consistently, publicly available H3K27ac HiChIP data in HT29 [68] revealed loops linking this CORE to the *USP7* promoter, providing independent contact-based supportive evidence for a regulatory connection between CORE and *USP7* gene expression (Extended Data Fig. 8a). These results exemplify that enCORE can move from disease-state stratification to actionable hypotheses by linking enhancer clusters, candidate genes, and regulatory evidence. In conclusion, we demonstrated that enCORE effectively elucidates CRC-dependent epigenetic programs in epithelial cells and identifies the potential therapeutic targets.

Discussion

This study introduces enCORE, a network-based framework that infers highly interactive enhancer clusters from scATAC-seq data by leveraging enhancer-enhancer interaction networks. Unlike element-centric workflows, enCORE yields a domain-scale representation of regulatory architecture in which coordinated chromatin accessibility, interaction structure, and TF occupancy are jointly considered. Across distinct biological contexts, enCORE enables the inference of highly interactive enhancer clusters and links chromatin organization to cell identity and disease.

The widespread adoption of single-cell epigenomics has made apparent a key limitation of population-averaged regulatory domains derived from bulk-level profiles. Such annotations can blur cell-state heterogeneity and mask epigenetic programs present in subpopulations. This motivates the need to study large cis-regulatory domains directly in the single-cell setting, where domain-scale organization can be resolved in a state-aware manner rather than fragmented into isolated peaks. enCORE addresses this gap by stratifying heterogeneous populations and summarizing context-dependent regulatory architecture at the level of highly interactive enhancer clusters. For example, in PBMC and BMMC, CORE profiles recapitulated hematopoietic lineage structure and recovered canonical lineage-defining regulators, supporting that domain-scale summaries can remain interpretable even under single-cell sparsity. As atlas-level single-cell datasets accumulate, domain-level representations may

become increasingly valuable for investigating state-resolved epigenetic programs.

From the perspective of large cis-regulatory domains, the defining feature of highly interactive enhancer clusters is not any single element, but the cooperative organization of individual enhancers into a coordinated regulatory assembly stabilized by spatial proximity. CORE encodes this domain concept by prioritizing ensembles of enhancers that form locally coherent structures in enhancer-enhancer interaction networks, thereby moving beyond element-centric signals. This conceptualization implies that large enhancer domains can be described not only by broad activity marks, but also by network topology that reflects how constituent elements are wired together. In this context, CORE provides a topology-aware domain representation that enables tractable interrogation of how regulatory assemblies are maintained or rewired during cell state transitions at the single-cell level.

CORE and SE share conceptual similarity as enhancer clusters, yet they prioritize different aspects of regulatory architecture. Our analyses indicate that CORE contains a higher fraction of poised/primed enhancer constituents than SE. Moreover, whereas SEs are more closely aligned with degree-ranked hubs, COREs tend to favor regions with higher clustering coefficients in the enhancer-enhancer interaction networks. In CRC, CORE preferentially colocalized with cancer-elevated H3K27ac programs and prioritized key cancer-associated regulators that were partially overlapping yet distinct from SE-associated sets, suggesting that CORE can delineate context-dependent regulatory programs that may be underrepresented in SE-derived annotations. Together, these findings position enCORE as a complementary strategy to SE for interpreting context-dependent epigenetic regulation at the single-cell level.

Our work has several limitations. Although enCORE was designed around three major characteristics of highly interactive enhancer clusters, the current version does not indicate which characteristics are most influential in a given context, nor does it clarify which characteristics are most determinative in its assignment as CORE for a given region. To address this interpretability gap, future work will incorporate an explainable modeling using generalized linear models (GLM) or XGBoost to quantify per-context feature importance and to assign per-region attributions. Further limitations include reliance on publicly available datasets, the absence of validation in other organisms (e.g. mouse), and that all evaluations were conducted *in silico*; additional experimental studies would be required to establish causal roles in gene regulation.

Collectively, enCORE provides a flexible framework for inferring highly interactive enhancer clusters and for decoding domain-level epigenetic programs across various biological contexts, including lineage-biased priming or disease-dependent reprogramming.

Methods

enCORE framework

enCORE framework is a workflow based on the following procedures: (1) Enhancer candidates extraction, (2) Initial decomposed components (enhancer clusters) inference, (3) Link loss correction, (4) Automatic distance threshold calculation, (5) Sparsity correction (rescue steps for enhancer pairs with high cosine similarity and lost peaks with high accessibility), (6) Pseudo-directed graph construction (with network node and edge weight assignment) (7) personalized PageRank-based ranking for enhancer clusters, (8) Distilling raw CORE ("potential" option) or filtered CORE ("active" option) after iterative proximal enhancer filtering procedure.

There are two options for analysis with the enCORE framework: potential and active. The "potential" option is the original one for the enCORE algorithm. Given that chromatin interactions can precede transcriptional activation [69, 70], poised enhancers exhibit high accessibility [71], polycomb domain near lineage-specifying developmental regulators also occur within TADs, as do SE domains [72], and super-silencers are present [5, 6], the potential option may, to some extent, inherently reflect not only the *current active* state but also potential *future-responsive poised or primed* epigenetic or transcriptional activation states. These characteristics may be useful for predicting transcriptional changes associated with disease, development, or environmental stimuli; however, from the perspective of investigating *current* transcriptional regulation, they could pose potential obstacles or limitations.

Hence, the "active" option was designed to emphasize the *current* state under a specific biological context. In the active option, the *iterative proximal enhancer clusters filtering* step, filtered by promoter accessibility, was applied to the original CORE from the potential option.

gABC scoring & Extracting initial enhancer clusters.

Peaks with a co-accessibility (correlation coefficient) of 0.2 or higher were defined as initial enhancer candidates if they were located more than 2kb away from the transcription start site (TSS) of UCSC known genes and showed co-accessibility with at least one promoter. The normalized accessibility

(CPM) of these initial enhancer candidates was used as input to the gABC scoring procedure implemented in STARE [73]. Peaks with a gABC score greater than 0.01 were redefined as filtered enhancer candidates annotated with gABC scores for multiple genes. For each cell population, we calculated pairwise cosine similarity between co-accessible filtered enhancer candidates based on the gABC score. Subsequently, co-accessibility network edges with a cosine similarity below the specific cut-off (default: 0.4) were removed using the decompose function from the igraph package (v1.5.1). The remaining decomposed components (disconnected clusters in the network) could be defined.

Link loss correction.

We assumed that some co-accessibility network edges may be lost because of the sparsity of scATAC-seq data. This can make more decomposed components compared with the expected number of enhancer clusters. Thus, pairwise cosine similarities between any two components within 500kb were recalculated using the summed values of gABC scores of enhancers in each component to correct the loss of network edges. The farthest two components with a cosine similarity greater than the specific cut-off (default: 0.4) and all intermediate components between them were merged into a single component. The final components obtained after the merging process were defined as LLS-corrected enhancer clusters.

Distance threshold optimization & Sparsity correction.

Details of the distance threshold optimization process are as follows. LLS-corrected enhancer clusters are likely to span excessively large genomic distances. Based on the assumption that the likelihood of functionally effective three-dimensional (3D) chromatin interactions decreases with increasing one-dimensional (1D) genomic distance, we automatically determined the effective distance threshold for interactions and re-segmented each LLS-corrected enhancer cluster accordingly.

To infer the distance threshold, we employed a strategy that maximizes both the following two conditional probabilities for the genomic distance between the two closest enhancer candidates in the intra-cluster (within a specific enhancer cluster) and inter-cluster (between two enhancer clusters)

contexts.

$$a = P(D < D_t|intra), \qquad b = P(D \geq D_t|inter) \ (1)$$

where $P(D < D_t|intra)$ is the conditional probability that the intra-cluster distance D is smaller than a specific threshold D_t, and $P(D \geq D_t|inter)$ is the conditional probability that the inter-cluster distance D is bigger than a specific threshold D_t. We reformulated the problem as maximizing the harmonic mean of these two conditional probabilities. The harmonic mean is defined as:

$$\mathrm{H} = \frac{2ab}{a+b} \ (2)$$

Given that both a and b are positive and their product $ab = constant$, the harmonic mean H reaches its maximum when $(a+b)$ is minimized.

Considering the following equation:

$$a + b = \sqrt{(a-b)^2 + 4ab} \ (3)$$

Under $ab = constant$, H is maximized when $a - b = 0$, i.e., when $a = b$. We assumed that intra-cluster and inter-cluster distance distributions are well separated, with substantially different means and minimal overlap.

Under this assumption, in an interval close to the point where H is maximized (D_H), an increase in one conditional probability necessarily entails a decrease in the other. Let $f_a(D)$ and $f_b(D)$ denote the corresponding probability density functions. Then,

$$a' = f_a(D), \qquad b' = -f_b(D) \ (4)$$

Because the overlap between the intra-cluster and inter-cluster distance distributions is assumed to be minimal, for an arbitrary sufficiently small positive constant $\varepsilon \ (0 < \varepsilon \ll 1)$, there exists an interval around D_H in which

$$f_a(D) \leq \varepsilon, \qquad f_b(D) \leq \varepsilon \ (5)$$

Meanwhile, since $0 \leq a \leq 1$ and $0 \leq b \leq 1$,

$$|(ab)'| = |f_a(D)b - f_b(D)a| = f_a(D)b + f_b(D)a \leq f_a(D) + f_b(D) \ (6)$$

Around D_H,

$$|(ab)'| \leq 2\varepsilon \ (7)$$

Thus, in the interval near D_H, $(ab)' \approx 0$, so ab can be approximated as constant. Based on this estimation, the distance threshold can be inferred as a genomic distance value when the difference between the two conditional probabilities $(a - b)$ is very close to zero. We defined the threshold as the distance at which $(a - b)$ first becomes greater than zero, for all bins generated by dividing the 500bp to 100kb range into 500bp intervals.

The sparsity correction procedure consists of two steps:

First, we additionally merged two nearest enhancer candidates into the same cluster if they were included as enhancer candidates after gABC scoring, assigned to different enhancer clusters after distance threshold optimization, yet showed a very high cosine similarity, suggesting that they should reasonably belong to the same enhancer cluster. This approach is justified by the assumption that some enhancers are filtered during the gABC scoring process, and the inherent sparsity of scATAC-seq data can lead to this discrepancy.

Second, for peaks that were not included as enhancer candidates after gABC scoring but existed in the initial set of enhancer candidates before scoring, we reconsidered their inclusion if they (1) had a co-accessibility (correlation coefficient) greater than a specific threshold (default: 0.49) with at least one enhancer candidate and (2) showed very high accessibility. The default accessibility threshold was defined as the accessibility value of the top 0.5% of enhancer candidates after gABC scoring. Newly added enhancer candidates were assigned to the enhancer cluster of the linked enhancer candidate that exhibited the highest value of assignment score. Assignment score is defined as follows:

$$(Assignment\ score)_{po} = \frac{corr(e_p, e_o) \times acc(e_o)}{dist(e_p, e_o)} \ (8)$$

where $corr(e_p, e_o)$ is the correlation coefficient between the newly added enhancer candidate e_p and the original enhancer candidate e_o, $acc(e_o)$ is the accessibility of e_o, and $dist(e_p, e_o)$ is the genomic distance between e_p and e_o.

Pseudo-directed network construction.

We assumed that the strength of the interaction between two nodes (two enhancer candidates) in the co-accessibility network is proportional to the co-accessibility (correlation coefficient), the inverse of the genomic distance, and the product of their accessibility values.

$$(interaction\ strength)_{ij} = \frac{corr(e_i, e_j) \times acc(e_i) \times acc(e_j)}{dist(e_i, e_j)} \quad (9)$$

where $corr(e_i, e_j)$ is the correlation coefficient between two different nodes e_i and e_j, $acc(e_i)$ is the accessibility of e_i, $acc(e_j)$ is the accessibility of e_j, and $dist(e_i, e_j)$ is the genomic distance between e_i and e_j. Based on the co-accessibility network, we constructed a network with the above interaction strength as the network edge weight.

Inspired by the concept of electronegativity in chemistry, we assumed that the interaction strength assigned to each edge is asymmetrically distributed between two connected nodes in proportion to the total accessibility of the enhancer cluster to which each node belongs. To incorporate this weight into the calculation of node centrality (personalized PageRank ranking procedure), we transformed the constructed network from an undirected graph to a directed graph. Specifically, for each edge between node e_i and node e_j, we redistributed the previously calculated interaction strength weights (as shown in Equation (6)) into two directed edges, one from e_i to e_j and the other from e_j to e_i, based on the relative total accessibility of their corresponding enhancer clusters (Figure 1b). The network edge weight for an edge from e_i to e_j is as follows.

$$(network\ edge\ weight)_{ij} = \frac{acc(c_j)}{(acc(c_i) + acc(c_j))} \cdot (interaction\ strength)_{ij} \quad (10)$$

where $acc(c_i)$ is the total summed accessibility of the enhancer cluster c_i corresponding to node $e_{i,}$ and $acc(c_j)$ is the total summed accessibility of the enhancer cluster c_j corresponding to e_j. After assigning edge weights in the network, we further assigned node weights for the personalized PageRank calculation. The node weight is defined as follows:

$$(network\ node\ weight)_i = MinMax(acc(e_i)) \cdot MinMax(w_{TF}(e_i)) \quad (11)$$

where $MinMax(acc(e_i))$ is the value set of $MinMax$-scaled accessibility of node e_i across all nodes

within the network, and $MinMax(w_{TF}(e_i))$ is the value set of $MinMax$-scaled TF weight of e_i across all nodes within the network. $MinMax$-scaling was performed using the scales package (v1.3.0).

TF weight calculation for each node was performed as follows: First, we applied motifmatchr (v1.20.0) to identify the presence of CISBP TF motifs in each final enhancer candidate (enhancer constituent). Subsequently, motif deviation scores (z-scores) computed per cell using ArchR (v1.0.2) were aggregated by calculating the median across all cells within the target cell population. The median z-score of each TF was defined as the motif enrichment weight (per cell population & per TF). TF motifs with motif enrichment weight below a pre-defined threshold (default: 0.675) were assigned a weight of 0. Finally, for each enhancer constituent, the motif enrichment weights for all present TF motifs were summed to obtain the node-specific TF weight.

personalized PageRank ranking.

For the constructed pseudo-directed network, we calculated personalized PageRank values using both node and edge weights using the igraph package (v1.5.1). Then, the PageRank values assigned to each node were aggregated per enhancer cluster, and the resulting sum was defined as the PageRank sum representing the relative importance of each enhancer cluster. For normalization, we scaled the maximum PageRank sum to 100 and ranked it in ascending order based on the scaled PageRank sum of the enhancer clusters.

In this context, the kneed package (v1.0.0) was adopted to determine the threshold of the scaled PageRank sum value corresponding to the elbow point of the ranked distribution. The region of the enhancer cluster where the scaled PageRank sum exceeds this threshold was defined as the Clustered Open Regulatory Elements (CORE). The initial sensitivity parameter was set to 2 to infer the elbow point. The threshold was accepted if the number of CORE regions identified using that elbow point did not exceed the maximum allowed number of CORE regions (default: 0.1*26000/1.31). Otherwise, the sensitivity parameter was repeatedly increased by 1, and the elbow points were recalculated until the number of CORE regions fell below the maximum limit. In some cases, constituents within the defined CORE may exhibit asymmetric chromatin accessibility, where regulatory activity is concentrated in only a subset of regions that cooperatively regulate similar gene expression programs. Consequently, this

can result in overestimating the spatial and functional span of CORE. To address this problem, we excluded constituents with accessibility values below a specific threshold (default: 5) from the final CORE definition.

Iterative proximal enhancer clusters filtering.

We annotated CORE regions to their nearest genes using ChIPseeker (v1.34.1). The promoter accessibility of each annotated gene was defined as the sum of the normalized accessibility values of all peaks located within the promoter region (within 2kb from the TSS). Genes with promoter accessibility not exceeding 12.5 were considered to have inactive promoters. Subsequently, the genomic regions within a specific distance (default: 7.5kb upstream and downstream from the TSS of inactive promoters) were defined as inactive proximal enhancer regions. Using BEDTools (v2.31.0) subtract function, we subtracted inactive proximal enhancer regions from the raw CORE regions, resulting in a preliminary filtered set of CORE regions. The initially filtered CORE regions were then re-annotated to the nearest genes. If the initially filtered CORE region was re-annotated as a gene with a TSS overlapping the inactive proximal enhancer region and was truncated in the previous first filtering step, this CORE region was subsequently eliminated. After this additional filtering step, the remaining CORE regions were defined as the final output of the active option.

scATAC-seq & scCUT&Tag-seq processing

Data retrieved from publicly available scATAC-seq datasets: 10k PBMC from 10X Genomics, BMMC (GSM4138889) [17], and CRC (GSE201349) [18]. ArchR (v1.0.2) was used for processing and analyzing scATAC-seq data. Low-quality cells were filtered based on two quality measures: Cells with nFrags>1000 and TSSenrichment>4 were retained. Doublets were detected and removed by using the ArchR filterDoublet function (filterRatio=1.0). The remaining cells were normalized and subjected to downstream analysis using the ArchR pipeline. Dimensionality reduction was performed using the Iterative Latent Semantic Indexing (IterativeLSI) method. Harmony method was applied to the reduced dimensions in order to remove unwanted non-biological variations. UMAP visualization was conducted based on the Harmony-corrected reduced dimensions. Imputed gene scores were calculated using

MAGIC (v3.0.0).

With the exception of PBMC data, cell types were annotated based on MAGIC-imputed gene scores of canonical marker genes. For PBMC, to perform cell type annotation, the label transfer method was applied using Signac (v1.14.0) before ArchR processing. Normalized ATAC signals (by ReadsInTSS metric) were used to generate the genome browser track visualization result. Co-accessibility information was extracted by using the ArchR addCoAccessibility function with these parameters: corCutOff=0.2 and maxDist=500000.

For scCUT&Tag-seq analysis, we used publicly available data (GSE195725) [28] for PBMC. The .rds files from GSE195725 were processed using Signac. Cell types were inferred based on clustering information and pre-annotated labels. Peaks were called for each cell type, and regions within 2kb of UCSC known genes' TSS were excluded. The remaining peaks were stitched using a maximum distance threshold of 12.5kb. Finally, super-enhancers (SE) were identified using the ROSE algorithm.

ChIP-seq processing

For validation of PBMC data, ChIP-seq datasets for H3K27ac, H3K4me1, and H3K27me3 were mainly collected from the ENCODE database [21], except for CD16+ Monocytes. In the case of CD16+ Monocytes, H3K27ac ChIP-seq data were obtained from the FANTOM database (GSE40502) [74]. Raw FASTQ files were processed as follows: adapter trimming with Skewer (v0.2.2), hg19 genome alignment with Bowtie2 (v2.5.1), de-duplication with the PICARD (v2.27.4) MarkDuplicates command, and blacklist filtering with BEDTools (v2.31.0) intersect function. The blacklist-filtered BAM files were then used to construct Tag Directories with the HOMER (v4.11.1) makeTagDirectory command. For track visualization for CD14+ Monocytes, we used CD14+CD16- Monocytes data from GSE40502 and applied the same processing pipeline as described above. BedGraph files were generated with HOMER makeUCSCfiles function. Track visualization was performed using WashU Epigenome Browser [75].

For validation of CRC data, two publicly available datasets were used: H3K27ac data from GSE156613 [52] and H3K27ac & H3K4me1 data from GSE136889 [53]. To infer SE regions of CRC, 15 CRC samples were sub-sampled from each dataset (Supplementary Table 8). Additionally, 15 normal colon samples were chosen from GSE136889 for comparative analysis (Supplementary Table 8). Adapter

trimming was performed using Skewer. Trimmed reads were aligned to the hg38 genome using Bowtie2. Duplicated reads were removed using the PICARD MarkDuplicates command. Blacklist regions were filtered using BEDTools intersect function. The blacklist-filtered BAM files were then used to construct Tag Directories with the HOMER makeTagDirectory command. For classification performance comparison between CORE and SE regions in CRC and normal colon samples, raw read count matrices were extracted using DiffBind. Normalization was performed using DESeq2, with size factors calculated as library size divided by median library size across samples. For downstream analysis, rlog-transformed counts were used for Non-Negative Factorization (NMF; using NMF R package v0.28; random.seed=1-100), and normalized counts were used for XGBoost classification (after hyperparameter tuning via Grid Search with cv=5; random.seed=1-100 in iterative random holdout procedure with 0.75:0.25 train test split for test set validation).

SE calling was performed using HOMER findPeaks with the parameters -style super -L 0. Regions within 2kb of UCSC known genes' TSS were excluded from SE identification using the -excludePeaks option.

Bulk RNA-seq processing

To investigate cell type-specific transcriptional regulation of CORE-associated and SE-associated genes in PBMC, we utilized the Monaco PBMC RNA-seq dataset and the HPA PBMC RNA-seq dataset from the Human Protein Atlas (HPA; https://www.proteinatlas.org/) [76]. For gene expression values, we used normalized TPM values pre-calculated by the Human Protein Atlas. To quantify the contribution of CORE and SE to cell type-specific transcriptional regulation, we calculated the mean log2 fold changes (Specificity score) of normalized TPM values between CORE-associated genes vs non-CORE-associated genes or SE-associated genes vs TE-associated genes for each cell type based on the Monaco dataset.

$$(Specificity\ score)_X = log_2\left(\frac{1 + mean\left(nTPM_{X,gene_X}\right)}{1 + mean\left(nTPM_{X^c,gene_{X^c}}\right)}\right) \quad (12)$$

where $mean\left(nTPM_{X,gene_X}\right)$ is the average value for normalized TPM of genes associated with region

set from samples corresponding to cell type X, and $mean\left(nTPM_{X^c,gene_{X^c}}\right)$ is the average value for normalized TPM of genes associated with region set from samples corresponding to other cell types X^c.

To measure the overlap between differentially expressed genes (DEGs), CORE-associated genes, and SE-associated genes in CLL vs Normal B cells, we utilized the publicly available dataset (GSE66117) [77]. Raw read counts were subjected to RLE normalization using DESeq2 (v1.38.3). DEGs were identified as DESeq2 by applying a cut-off of |log2FoldChange|>1 & q-value < 0.05. The log2 fold change values were shrunk using the apeglm package (v1.20.0). The degree of overlap between DEGs and CORE/SE-associated genes was quantified using the following formula:

$$(Degree\ of\ overlap)_R = \frac{N(gene_R\ \cap\ DEGs)}{N_{DEGs}} \quad (13)$$

where $N(gene_R\ \cap\ DEGs)$ is the number of overlaps between DEGs and genes associated with region R, and N_{DEGs} is the number of DEGs.

Peak annotation & GO enrichment analysis

GREAT [78] was used to identify CORE-associated and SE-associated genes. For PBMC and BMMC, peak-to-gene annotation was performed using a single nearest gene within 500kb of TSS as a metric. In contrast, for CRC, we used the two nearest genes within the same distance threshold. In addition, GO enrichment analysis for CD14+ and CD16+ Monocytes of PBMC were also performed using GREAT.

Chromatin interaction enrichment analysis

For H3K27ac HiChIP data analysis, FitHiChIP 5k loop information for GM12878, THP-1, and Th17 was obtained from HiChIPdb [22]. Using loop information, we constructed the chromatin interaction network with the igraph package (v1.5.1) and calculated weighted degree (strength) for each node. Normalized count values were used as network edge weights.

For Hi-C data analysis, loop information (.bedpe files with hiccups) for B cells (ENCSR847RHU) and

CD14+ Monocytes (ENCSR236EYO) was obtained from the ENCODE database [21]. Loops with FDR > 0.1 were removed (observed/expected ratio-based FDR). After that, only the loops in which the two interaction regions overlapped with the H3K4me1 peaks of the corresponding cell type (CD19+ B cells: ENCSR214VUB, CD14+ Monocytes: ENCSR400VWA) were retained. We then constructed the enhancer-enhancer chromatin interaction network using igraph and calculated weighted degree (strength) for each node. Network edge weights were defined as observed-minus-expected counts to quantify contact signals above the local expected background. This metric was used to preserve differences in interaction magnitude, as constituent-level interaction strength derived from ratio-based measures can be attenuated in interaction-dense neighborhoods with elevated local expected counts. As the Hi-C data were provided in hg38, CORE and SE coordinates originally defined in hg19 were converted to hg38 using UCSC LiftOver [79].

To characterize the epigenetic states of enhancer constituents within CORE and SE, we categorized H3K4me1-marked enhancer peaks into three state categories: active, poised, and primed enhancers. The detailed procedure is as follows: Among the ChIP-seq datasets described in the *ChIP-seq processing* section above, we used the H3K27ac for CD19+ B cells (ENCSR191ZQT) and CD14+ Monocytes (ENCSR012PII), and the H3K27me3 for CD19+ B cells (ENCSR404MOX) and CD14+ Monocytes (ENCSR080XUB) from the ENCODE database [21]. Read count matrices were generated for H3K27ac and H3K27me3 at pre-defined H3K4me1 peaks using the DiffBind package (v3.8.4). The data was normalized with CPM (Counts per Million). H3K4me1 peaks were classified into three state categories based on the log2 fold change between H3K27ac and H3K27me3 signals: peaks with a value greater than 1 were classified as active, those with a value less than -1 as poised, and the remaining peaks as primed.

GWAS enrichment analysis

For GWAS enrichment analysis, in the present study, we collected GWAS variants associated with immune-related diseases and non-immune-related traits through two approaches. First, we utilized the fine-mapped GWAS variants (95% credible set from FINEMAP metric) curated from several GWAS cohorts via CausalDB [19]. Second, to facilitate direct comparison with SE, we used the Young Lab-

curated version of GWAS variants (Hnisz *et al.*) that first reported the correlation between SE and disease-related GWAS enrichment [2]. To measure GWAS enrichment for CORE and SE, we defined the enrichment score as follows:

$$ES_P = \frac{N_{Obs}(P \cap R)}{N_{Exp}(P \cap R)} \quad (14)$$

where $N_{Obs}(P \cap R)$ is the observed number of GWAS variants included in the GWAS phenotypic trait P within region set R, and $N_{Exp}(P \cap R)$ is the average number of GWAS variants included in the GWAS phenotypic trait P within a randomly permuted region set derived from region set R with 1000 iterations.

To measure cell type specificity for fine-mapped CAD-risk loci enrichment, the specificity index was defined as follows.

$$(Specificity\ index)_X = \frac{(ES_{CAD,X})}{\max\,(ES_{CAD})} \quad (15)$$

where $ES_{CAD,X}$ is the GWAS enrichment score for fine-mapped CAD risk loci in cell type X, and $\max\,(ES_{CAD})$ is the maximum value of the GWAS enrichment scores from all cell types, including X.

TF motif enrichment analysis

To perform both known motif enrichment and *de novo* motif enrichment analyses, the HOMER (v4.11.1) findMotifsGenome.pl function was used with the -size given & -mask parameters. For PBMC data, only the known motif enrichment analysis results were adopted. Calculation of the motif enrichment score was performed using the modified MinMax scaling method as follows.

$$(Motif\ enrichment\ score)_{MX} = 6 \times \frac{V_{MX} - \min\,(V_M)}{\max(V_M) - \min\,(V_M)} \quad (16)$$

where V_{MX} is the statistical significance (-log10(P-value)) for TF motif M within CORE regions from any cell type X, $\max(V_M)$ is the maximum V_{MX} across all cell types, and $\min\,(V_M)$ is the minimum V_{MX} across all cell types. Modified $MinMax$-scaling was performed using the scales package (v1.3.0).

Only significant motifs with p-value < 1e-04 & q-value < 0.05 were included in the known motif

enrichment analysis. For CRC, the results from the *de novo* motif enrichment analysis were used. Significant motifs for *de novo* motif enrichment analysis were defined as those not marked by possible false positives.

Similarity network construction for bone marrow and peripheral blood hematopoietic cells

To incorporate the CORE regions derived from each BM cell type, we first employed the BEDTools (v2.31.0) sort and merge function. We used the BEDTools intersect function to determine whether the merged regions overlapped with the CORE regions of each cell type. Overlaps were binarized as 1 (overlap) or 0 (no overlap), resulting in a binary region x cell type matrix format. Based on the binary matrix, we calculated the pairwise cosine similarity between cell types using the coop package (v0.6-3). We then constructed the network with the igraph package (v1.5.1) using the cosine similarity matrix. Edges with cosine similarity < 0.55 were removed to refine the network. Network visualization was performed using the ggraph package (v2.2.1). For network construction on BM-to-PB maturation ordering, the same procedure was applied to PB cell types. Pairwise cosine similarities were calculated between BM and PB cell types. Then, edges with cosine similarity < 0.55 were removed to refine the BM-to-PB bipartite network.

Single-cell Multiome velocity analysis

We analyzed the single cell multiome dataset for CD34+ hematopoietic cells provided in the original MultiVelo paper and tutorial [46]. Cell types were inferred based on pre-annotated labels. For Multiome velocity estimation, we used the following parameters: n_pc=10, n_neighbors=20, weight_c=0.75.

Enrichment analysis of H3K27ac differential peaks between macrophages and monocytes

Raw FASTQ files for H3K27ac profiles of macrophages and monocytes were retrieved from publicly available datasets (GSE31621) [80]. Then, using the same pipeline as above *ChIP-seq processing* for PBMC, reads were aligned to the hg19 reference genome. Both blacklist-filtered BAM files and BedGraph files were generated. Using DiffBind, raw read count matrices were extracted. CPM (Counts

per million) normalization was applied to the raw read count matrices. Differential peaks between macrophages and monocytes were obtained using |log2FoldChange| > 3 & q-value < 0.05 as a cut-off value. The Cochran-Mantel-Haenszel (CMH) test was used for enrichment analysis. The detailed procedure is as follows: To compare CORE and SE enrichment, we used the regions where TE did not overlap with any CORE or SE as a control peak set. Peak widths for the three types of peak set (CORE, SE, and TE) were binned to quantile ranges. With the generated bin as strata, the CMH test was applied to contingency tables summarizing the overlap between differential peaks (macrophage-elevated, monocyte-elevated) and each peak set comparison (CORE vs TE, SE vs TE), yielding common odds ratios and p-values. The Breslow-Day (BD) test was used as a homogeneity test to determine the reliability of the calculated common odds ratios. For all comparisons, the common odds ratio was determined to be confident by satisfying p-value > 0.05 derived from the BD test.

TCGA survival analysis

Survival analysis for CORE-associated genes was performed using the Colon Adenocarcinoma (COAD) cancer subtype data from the TCGA database. Survival curves were produced using the survival package (v3.4-0). Statistical significance P values were computed using a log-rank test.

Enrichment analysis for CRC master TF candidates

Master TF candidates for CRC were collected from previous reports [66, 81-83]. The enrichment score was defined as the proportion of CRC master TF candidates that overlap with CORE- or SE-associated genes using the following formula:

$$(Proportion)_R = \frac{N(gene_R \cap CRC)}{N_{CRC}} \quad (17)$$

where $N(gene_R \cap CRC)$ is the number of overlaps between CRC master TF candidates and genes associated with region R, and N_{CRC} is the number of CRC master TF candidates.

Network-based analysis and Metastasis-related genes

Based on the co-accessibility network information of epithelial cells in the CRC scATAC-seq dataset [18], an undirected graph was constructed using the igraph package (v1.5.1), with correlation coefficients assigned as network edge weights. We then calculated weighted degree (strength) and local clustering coefficient (transitivity) for each node. Both metastasis-related TFs and target genes were obtained from the COAD-related single-cell RNA-seq and bulk RNA-seq datasets provided by metsDB [23]. Metastasis-related TFs were identified from the single-cell RNA-seq dataset. Metastasis-related target genes were defined as genes satisfying |logFC| > 2 & p_val_adj < 0.05 in the bulk RNA-seq dataset and predicted targets of the metastasis-related TFs in the single-cell RNA-seq dataset. The abundance of metastasis-related TFs and target genes was quantified using the following formula:

$$(Abundance)_R = \frac{N(gene_R \cap met)}{N_{met}} \quad (18)$$

where $N(gene_R \cap met)$ is the number of overlaps between metastasis-related TFs or target genes and genes associated with region R, and N_{met} is the number of metastasis-related TFs or target genes.

Additionally, to measure the enrichment of metastasis-related TFs and target genes in peaks with high weighted degree vs high local clustering coefficient (Top 10% H3K4me1 peaks ranked by clustering coefficient or degree), fold changes for the observed number of peak-annotated genes that overlap with metastasis-related genes (Obs) and the expected number of those genes (Exp) were calculated as follows.

$$\mathrm{Obs} = \frac{N(gene_l \cap met)}{N(gene_k \cap met)}, \mathrm{Exp} = \frac{N_l}{N_k} \quad (19)$$

where $gene_l$ is the gene set associated with peaks that have a high local clustering coefficient, $gene_k$ is the gene set associated with peaks that have a high weighted degree, $N(gene_l \cap met)$ is the number of overlaps between metastasis-related TFs or target genes and genes within $gene_l$, $N(gene_k \cap met)$ is the number of overlaps between metastasis-related TFs or target genes and genes within $gene_k$, N_l is the number of genes within $gene_l$, and N_k is the number of genes within $gene_k$.

***In silico* enhancer knockout analysis**

To perform *in silico* enhancer knockout analysis, we used AlphaGenome (v0.0.1) [67] with the Left colon ontology (UBERON0008971). We first selected candidate enhancer peaks from the CRC H3K27ac profiles [53] by extracting typical enhancer peaks that overlapped the USP7-proximal CORE. To minimize unwanted effects related to transcript structure (e.g., splicing), we restricted the analysis to peaks annotated as distal intergenic. The DNA sequence corresponding to the selected peak was retrieved using the getSeq function in the BSgenome package (v1.66.3) and used as the reference variant. Enhancer knockout was modeled by setting the alternative variant to a deletion (empty sequence), and the predict_variant function was used to predict the knockout-associated changes in the Left colon RNA-seq profile. The prediction interval was defined as a 1Mb window centered on the USP7 locus (SEQUENCE_LENGTH_1MB). To measure variant effects, variant scores were computed for each gene under the Left colon ontology using the tidy_scores function. As independent chromatin contact evidence, we incorporated publicly available H3K27ac HiChIP data from the HT29 cell line (GEO: GSM7117686) [68]. Intra-chromosomal loop coordinates provided in hg19 were converted to hg38 using UCSC LiftOver [79]. The lifted-over loops were visualized together with CRC & Normal H3K27ac ChIP-seq [53] signal tracks using the WashU Epigenome Browser [75].

Statistics and reproducibility

No statistical method was adopted to predetermine sample size. The experiments in the present study were not randomized, as all datasets are publicly available from observational studies. In case of the XGBoost classification, the training set and test set were randomly selected, and this analysis was repeated 100 times. The investigators were not blinded to allocation during experiments and outcome assessment. For hypothesis testing in a single group, Statistical significance P values were computed using a two-sided Wilcoxon signed-rank, Wilcoxon rank sum test. For multiple group comparison, P values were calculated using Dunn's test as a post-hoc test after the Kruskal-Wallis test. Multiple hypothesis correction was performed by the Bonferroni correction method.

Data availability

The publicly available datasets were downloaded via Gene Expression Omnibus; scATAC-seq (GSM4138889, GSE201349), scCUT&Tag-Pro (GSE195725), ChIP-seq (GSE40502, GSE156613, GSE136889, GSE31621), RNA-seq (GSE66117), and HiChIP (GSM7117686). scATAC-seq dataset for 10X PBMC 10k was retrieved from a web repository (https://cf.10xgenomics.com/samples/cell-atac/1.0.1/atac_v1_pbmc_10k/atac_v1_pbmc_10k_fragments.tsv.gz). Super-enhancer and typical enhancers for PBMC were downloaded from SEdb [20]. The Monaco PBMC RNA-seq dataset and the HPA PBMC RNA-seq dataset were obtained from the Human Protein Atlas (HPA; https://www.proteinatlas.org/). Hi-C loop information and ChIP-seq for H3K27ac, H3K27me3, and H3K4me1 (CD4+ T cells, CD8+ T cells, NK cells, B cells, and CD14+ Monocytes) were collected from the ENCODE database [21]. HiChIP for H3K27ac (GM12878, THP-1, and Th17) were obtained from HiChIPdb [22]. This work used fine-mapped and curated GWAS variants for immune-related diseases and non-immune traits. Fine-mapped GWAS variants used in this paper were downloaded from CausalDB [19]. Young Lab-curated version of GWAS variants used in this paper are available through the original publication with PMID: 24119843. Single cell multiome dataset for CD34+ hematopoietic cells are available in the MultiVelo tutorial (https://github.com/welch-lab/MultiVelo). Metastasis-related TFs and target genes were obtained from metsDB [23].

Code availability

enCORE is available on GitHub at https://github.com/R-Krait/enCORE. Code for the PBMC 3k demo dataset can be found at https://figshare.com/articles/dataset/Demo_PBMC_3k_enCORE_/31577779.

Author information

Authors and Affiliations

Department of Biological Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan, 44919, South Korea

Seonjun Park, Sungkook Ma, Wonhyo Lee & Sung Ho Park

Contributions

S.P. and S.H.P. conceptualized the study. S.P. contributed to the design and implementation of the method and performed all computational analyses. S.P., S.M., W.L., and S.H.P. contributed to the interpretation of the results. S.P. and S.H.P. wrote the original draft, and all authors contributed to manuscript review and editing. All authors read and approved the final version of the manuscript.

Corresponding author

Correspondence to Sung Ho Park.

References

1. Whyte, Warren A., et al., *Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes.* Cell, 2013. **153**(2): p. 307–319.
2. Hnisz, D., et al., *Super-Enhancers in the Control of Cell Identity and Disease.* Cell, 2013. **155**(4): p. 934–947.
3. Parker, S.C.J., et al., *Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants.* Proceedings of the National Academy of Sciences, 2013. **110**(44): p. 17921–17926.
4. Benayoun, B.A., et al., *H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency.* Cell, 2015. **163**(5): p. 1281–1286.
5. Zhang, Y., et al., *Super-silencer perturbation by EZH2 and REST inhibition leads to large loss of chromatin interactions and reduction in cancer growth.* Nature Structural & Molecular Biology, 2025. **32**(1): p. 137–149.
6. Huang, D., et al., *Super-silencers are crucial for development and carcinogenesis in B cells.* Nature Communications, 2025. **16**(1): p. 8395.
7. Guelen, L., et al., *Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions.* Nature, 2008. **453**(7197): p. 948–951.
8. Madsen, J.G.S., et al., *Highly interconnected enhancer communities control lineage-determining genes in human mesenchymal stem cells.* Nature Genetics, 2020. **52**(11): p. 1227–1238.
9. Pott, S. and J.D. Lieb, *What are super-enhancers?* Nature Genetics, 2015. **47**(1): p. 8–12.
10. Kundaje, A., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317–330.
11. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**(7414): p. 75–82.
12. Ing-Simmons, E., et al., *Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin.* Genome Res, 2015. **25**(4): p. 504–13.
13. Buenrostro, J.D., et al., *Single-cell chromatin accessibility reveals principles of regulatory variation.* Nature, 2015. **523**(7561): p. 486–490.
14. Shi, P., et al., *Fundamental and practical approaches for single-cell ATAC-seq analysis.* aBIOTECH, 2022. **3**(3): p. 212–223.
15. Zhang, K., et al., *A single-cell atlas of chromatin accessibility in the human genome.* Cell, 2021. **184**(24): p. 5985–6001.e19.
16. Satpathy, A.T., et al., *Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion.* Nature Biotechnology, 2019. **37**(8): p. 925–936.
17. Granja, J.M., et al., *Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia.* Nature Biotechnology, 2019. **37**(12): p. 1458–1465.

18. Becker, W.R., et al., *Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer.* Nature Genetics, 2022. **54**(7): p. 985–995.
19. Wang, J., et al., *CAUSALdb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies.* Nucleic Acids Research, 2020. **48**(D1): p. D807–D816.
20. Wang, Y., et al., *SEdb 2.0: a comprehensive super-enhancer database of human and mouse.* Nucleic Acids Research, 2023. **51**(D1): p. D280–D290.
21. Luo, Y., et al., *New developments on the Encyclopedia of DNA Elements (ENCODE) data portal.* Nucleic Acids Research, 2020. **48**(D1): p. D882–D889.
22. Zeng, W., et al., *HiChIPdb: a comprehensive database of HiChIP regulatory interactions.* Nucleic Acids Research, 2023. **51**(D1): p. D159–D166.
23. Wu, S., et al., *metsDB: a knowledgebase of cancer metastasis at bulk, single-cell and spatial levels.* Nucleic Acids Research, 2025. **53**(D1): p. D1427–D1434.
24. Liao, X., et al., *Krüppel-like factor 4 regulates macrophage polarization.* The Journal of Clinical Investigation, 2011. **121**(7): p. 2736–2749.
25. Guo, W., et al., *Notch signaling regulates macrophage-mediated inflammation in metabolic dysfunction-associated steatotic liver disease.* Immunity, 2024. **57**(10): p. 2310–2327.e6.
26. Yang, S., et al., *Decontamination of ambient RNA in single-cell RNA-seq with DecontX.* Genome Biology, 2020. **21**(1): p. 57.
27. Bartosovic, M., M. Kabbe, and G. Castelo-Branco, *Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues.* Nature Biotechnology, 2021. **39**(7): p. 825–835.
28. Zhang, B., et al., *Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro.* Nature Biotechnology, 2022. **40**(8): p. 1220–1230.
29. Schoenfelder, S., et al., *Divergent wiring of repressive and active chromatin interactions between mouse embryonic and trophoblast lineages.* Nature Communications, 2018. **9**(1): p. 4189.
30. Turner, C.A., Jr., D.H. Mack, and M.M. Davis, *Blimp-1, a novel zinc finger-containing protein that can drive the maturation of B lymphocytes into immunoglobulin-secreting cells.* Cell, 1994. **77**(2): p. 297–306.
31. Shapiro-Shelef, M., et al., *Blimp-1 Is Required for the Formation of Immunoglobulin Secreting Plasma Cells and Pre-Plasma Memory B Cells.* Immunity, 2003. **19**(4): p. 607–620.
32. Fillatreau, S., B. Manfroi, and T. Dörner, *Toll-like receptor signalling in B cells during systemic lupus erythematosus.* Nature Reviews Rheumatology, 2021. **17**(2): p. 98–108.
33. Seifert, M., et al., *Cellular origin and pathophysiology of chronic lymphocytic leukemia.* Journal of Experimental Medicine, 2012. **209**(12): p. 2183–2198.
34. Haskins, K. and M. McDuffie, *Acceleration of Diabetes in Young Nod Mice with a CD4+ Islet-specific T Cell Clone.* Science, 1990. **249**(4975): p. 1433–1436.

35. van der Ploeg, E.K., et al., *Type-2 CD8+ T-cell formation relies on interleukin-33 and is linked to asthma exacerbations.* Nature Communications, 2023. **14**(1): p. 5137.
36. Salou, M., et al., *Involvement of CD8+ T Cells in Multiple Sclerosis.* Frontiers in Immunology, 2015. **Volume 6 - 2015**.
37. Comi, G., et al., *Role of B Cells in Multiple Sclerosis and Related Disorders.* Annals of Neurology, 2021. **89**(1): p. 13–23.
38. Li, F. and H. Zhang, *Lysosomal Acid Lipase in Lipid Metabolism and Beyond.* Arteriosclerosis, Thrombosis, and Vascular Biology, 2019. **39**(5): p. 850–856.
39. Wild, P.S., et al., *A Genome-Wide Association Study Identifies LIPA as a Susceptibility Gene for Coronary Artery Disease.* Circulation: Cardiovascular Genetics, 2011. **4**(4): p. 403–412.
40. Farinello, D., et al., *A retinoic acid-dependent stroma-leukemia crosstalk promotes chronic lymphocytic leukemia progression.* Nature Communications, 2018. **9**(1): p. 1787.
41. Musumeci, A., et al., *What Makes a pDC: Recent Advances in Understanding Plasmacytoid DC Development and Heterogeneity.* Frontiers in Immunology, 2019. **Volume 10 - 2019**.
42. Kataoka, K., et al., *Evi1 is essential for hematopoietic stem cell self-renewal, and its expression marks hematopoietic cells with long-term multilineage repopulating activity.* Journal of Experimental Medicine, 2011. **208**(12): p. 2403–2416.
43. Gudmundsson, K.O., et al., *Prdm16 is a critical regulator of adult long-term hematopoietic stem cell quiescence.* Proceedings of the National Academy of Sciences, 2020. **117**(50): p. 31945–31953.
44. Germeshausen, M., et al., *MECOM-associated syndrome: a heterogeneous inherited bone marrow failure syndrome with amegakaryocytic thrombocytopenia.* Blood Advances, 2018. **2**(6): p. 586–596.
45. Jacobsen, S.E.W. and C. Nerlov, *Haematopoiesis in the era of advanced single-cell technologies.* Nature Cell Biology, 2019. **21**(1): p. 2–8.
46. Li, C., et al., *Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction.* Nature Biotechnology, 2023. **41**(3): p. 387–398.
47. Yagi, M., et al., *DC-STAMP is essential for cell–cell fusion in osteoclasts and foreign body giant cells.* Journal of Experimental Medicine, 2005. **202**(3): p. 345–351.
48. Garcia-Hernandez, M.d.l.L., et al., *Dendritic cell-specific transmembrane protein is required for synovitis and bone resorption in inflammatory arthritis.* Frontiers in Immunology, 2022. **Volume 13 - 2022**.
49. Cilenti, F., et al., *A PGE2-MEF2A axis enables context-dependent control of inflammatory gene expression.* Immunity, 2021. **54**(8): p. 1665–1682.e14.
50. Rauschmeier, R., et al., *Bhlhe40 and Bhlhe41 transcription factors regulate alveolar macrophage self-renewal and identity.* The EMBO Journal, 2019. **38**(19): p. e101233.
51. Saeed, S., et al., *Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity.* Science, 2014. **345**(6204): p. 1251086.
52. Lin, X., et al., *Cooperation of MLL1 and Jun in controlling H3K4me3 on enhancers in*

colorectal cancer. Genome Biology, 2023. **24**(1): p. 268.

53. Orouji, E., et al., *Chromatin state dynamics confers specific therapeutic strategies in enhancer subtypes of colorectal cancer.* Gut, 2022. **71**(5): p. 938.
54. Li, D., et al., *The Critical Role of Dysregulated FOXM1–PLAUR Signaling in Human Colon Cancer Progression and Metastasis.* Clinical Cancer Research, 2013. **19**(1): p. 62–72.
55. Zhang, X., et al., *A novel FOXM1 isoform, FOXM1D, promotes epithelial–mesenchymal transition and metastasis through ROCKs activation in colorectal cancer.* Oncogene, 2017. **36**(6): p. 807–819.
56. Napolitano, S., et al., *Targeting the EGFR signalling pathway in metastatic colorectal cancer.* The Lancet Gastroenterology & Hepatology, 2024. **9**(7): p. 664–676.
57. Stein, U., et al., *MACC1, a newly identified key regulator of HGF-MET signaling, predicts colon cancer metastasis.* Nature Medicine, 2009. **15**(1): p. 59–67.
58. Lemos, C., et al., *MACC1 Induces Tumor Progression in Transgenic Mice and Colorectal Cancer Patients via Increased Pluripotency Markers Nanog and Oct4.* Clinical Cancer Research, 2016. **22**(11): p. 2812–2824.
59. Zhou, X., et al., *Isthmin-1 promotes growth and progression of colorectal cancer through the interaction with EGFR and YBX-1.* Cancer Letters, 2024. **590**: p. 216868.
60. Yu, Y., et al., *PRDM15 interacts with DNA-PK-Ku complex to promote radioresistance in rectal cancer by facilitating DNA damage repair.* Cell Death & Disease, 2022. **13**(11): p. 978.
61. Chen, Z., et al., *Identification of cis-HOX-HOXC10 axis as a therapeutic target for colorectal tumor-initiating cells without APC mutations.* Cell Reports, 2021. **36**(4): p. 109431.
62. Borck, P.C., L.-W. Guo, and J. Plutzky, *BET Epigenetic Reader Proteins in Cardiovascular Transcriptional Programs.* Circulation Research, 2020. **126**(9): p. 1190–1208.
63. Lebeau, B., et al., *3D chromatin remodeling potentiates transcriptional programs driving cell invasion.* Proceedings of the National Academy of Sciences, 2022. **119**(36): p. e2203452119.
64. Shao, Z.-Y., et al., *The role of USP7-YY1 interaction in promoting colorectal cancer growth and metastasis.* Cell Death & Disease, 2024. **15**(5): p. 347.
65. Novellasdemunt, L., et al., *USP7 Is a Tumor-Specific WNT Activator for APC-Mutated Colorectal Cancer by Mediating β-Catenin Deubiquitination.* Cell Reports, 2017. **21**(3): p. 612–627.
66. Shin, D., et al., *Attractor Landscape Analysis Reveals a Reversion Switch in the Transition of Colorectal Tumorigenesis.* Advanced Science, 2025. **12**(8): p. 2412503.
67. Avsec, Ž., et al., *Advancing regulatory variant effect prediction with AlphaGenome.* Nature, 2026. **649**(8099): p. 1206–1218.
68. Mortenson, K.L., et al., *3D genomic analysis reveals novel enhancer-hijacking caused by complex structural alterations that drive oncogene overexpression.* Nature Communications, 2024. **15**(1): p. 6130.
69. Furlong, E.E.M. and M. Levine, *Developmental enhancers and chromosome topology.* Science, 2018. **361**(6409): p. 1341–1345.

70. Gorkin, David U., D. Leung, and B. Ren, *The 3D Genome in Transcriptional Regulation and Pluripotency.* Cell Stem Cell, 2014. **14**(6): p. 762–775.
71. Zentner, G.E., P.J. Tesar, and P.C. Scacheri, *Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions.* Genome Res, 2011. **21**(8): p. 1273–83.
72. Dowen, Jill M., et al., *Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes.* Cell, 2014. **159**(2): p. 374–387.
73. Hecker, D., et al., *The adapted Activity-By-Contact model for enhancer–gene assignment and its application to single-cell data.* Bioinformatics, 2023. **39**(2): p. btad062.
74. Schmidl, C., et al., *Transcription and enhancer profiling in human monocyte subsets.* Blood, 2014. **123**(17): p. e90–e99.
75. Seng, C., et al., *WashU Epigenome Browser update 2025.* Nucleic Acids Research, 2025. **53**(W1): p. W554–W561.
76. Uhlen, M., et al., *Towards a knowledge-based Human Protein Atlas.* Nature Biotechnology, 2010. **28**(12): p. 1248–1250.
77. Kushwaha, G., et al., *Hypomethylation coordinates antagonistically with hypermethylation in cancer development: a case study of leukemia.* Human Genomics, 2016. **10**(2): p. 18.
78. McLean, C.Y., et al., *GREAT improves functional interpretation of cis-regulatory regions.* Nature Biotechnology, 2010. **28**(5): p. 495–501.
79. Hinrichs, A.S., et al., *The UCSC Genome Browser Database: update 2006.* Nucleic Acids Research, 2006. **34**(suppl_1): p. D590–D598.
80. Pham, T.-H., et al., *Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states.* Blood, 2012. **119**(24): p. e161–e171.
81. Gong, J.-R., et al., *Control of Cellular Differentiation Trajectories for Cancer Reversion.* Advanced Science, 2025. **12**(3): p. 2402132.
82. Reddy, J., et al., *Predicting master transcription factors from pan-cancer expression data.* Science Advances. **7**(48): p. eabf6123.
83. Joanito, I., et al., *Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer.* Nature Genetics, 2022. **54**(7): p. 963–975.

Figure Legends

Fig.1: Overview of enCORE and its application to single-cell chromatin accessibility datasets

a The enCORE framework. Step 1: Extracting enhancer candidates using co-accessibility and gABC scoring. Step 2: Link loss correction and sparsity correction with distance threshold optimization. Step 3: Performing pseudo-directed graph construction on defined enhancer clusters. Step 4: Enhancer cluster ranking by personalized PageRank. Step 5: Two options for CORE calling; potential and active. **b** Detailed procedures for pseudo-directed graph construction. Converting an undirected co-accessibility network between enhancers to a directed graph. Left: Information to be used for network edge and node weight assignment. Middle: Assigning edge weights derived from accessibility of enhancers, co-accessibility coefficient, and contact frequency inferred by the inverse of distance. Right: Assigning node weights derived from the accessibility and TF information of enhancers with MinMax scaling. Node weights were used as the personalization vector for PageRank calculation. **c** Three single-cell ATAC datasets on which we applied enCORE to extract clustered open regulatory elements (CORE). UMAP embeddings of each dataset, colored and annotated by cell type. The average number of CORE per cell type of each dataset is represented below the UMAP embeddings.

Fig.2: enCORE recapitulates cell type-specific transcriptional regulation in PBMC

a Chow-Ruskey diagram of Non-CORE Active Enhancer (non-CORE)-associated genes and Clustered Open Regulatory Element (CORE)-associated genes in PBMC. B cells (red border), Monocytes (purple border), T cells (green border), NK cells (brown border). Monocytes represent the union of CORE-associated genes derived from CD14+ Monocytes and CD16+ Monocytes, while T cells represent the combined set of CORE-associated genes derived from CD4 Memory T cells and CD8 Effector T cells. The color of the borders around each intersection corresponds to the cell types whose genes overlap. The area of each intersection is proportional to the number of genes within the intersection. **b** Proportion of CORE constituents that overlap with H3K4me1 peaks, relative to the total number of enhancer candidates in each cell type. NK: NK cells, CD14_Mono: CD14+ Monocytes, B: B cells, CD8_Teff: CD8+ Effector T cells, CD4_Tmem: CD4+ Memory T cells. **c** Heatmap for cell type-specific master regulators. Black: gene present in the CORE-associated genes of the given cell type. White: absent. **d** Log2-

transformed fold changes of normalized TPM of genes in each cell type. Normalized TPM values were retrieved from the PBMC Monaco dataset in the Human Protein Atlas (HPA). Left: comparison between typical enhancer (TE)-associated genes and super-enhancer (SE)-associated genes. Right: comparison between non-CORE-associated genes and CORE-associated genes. P-value was calculated by a two-sided Wilcoxon signed-rank test. **e** Representative RNA expression for *GATA3*, *EBF1*, and *MAFB*. Normalized TPM values were retrieved from the HPA PBMC dataset. CD4 T Mem: CD4+ Memory T cells, CD8 T Mem: CD8+ Memory T cells, NK: NK cells, B mem: Memory B cells, B nai: naïve B cells, cDC: classical Dendritic cells, CD14 Mono: CD14+ Monocytes, CD16 Mono: CD16+ Monocytes. **f** Enrichment analysis of known TF motifs within CORE and non-CORE constituents across cell types. Colored by enrichment score, defined as modified MinMax-scaled statistical significance (-log10(P-value)).

Fig.3: CORE constituents show enriched chromatin interactions

a Log10-transformed weighted degrees of constituents within CORE and non-CORE. Data was retrieved from the H3K27ac HiChIP dataset for GM12878 cells. Left: CORE from the active option, Right: CORE from the potential option. **b** Log10-transformed weighted degrees of H3K4me1 peaks within CORE and non-CORE. Data was retrieved from the Hi-C dataset for CD19+ B cells. Only Hi-C interactions overlapping with H3K4me1 peaks are considered. CORE from the potential option. **c** Bar graphs showing the relative proportion of each enhancer category from CD19+ B cells within CORE and SE. active: active enhancers, poised: poised enhancers, primed: primed enhancers that do not fall into either the active or poised categories. **d** Log10-transformed weighted degrees of H3K4me1 peaks in each enhancer category from CD19+ B cells. Left: active enhancers, Middle: poised enhancers, Right: primed enhancers. Peak group marked by NON represents H3K4me1 peaks that do not fall into either CORE or SE. **e** Track visualization at the *PRDM1* locus with Hi-C interactions from CD19+ B cells. P-values were calculated to measure statistical significance (**a,b**, two-sided Wilcoxon rank sum test; **d**, Kruskal-Wallis test with Dunn's test; P-values were adjusted by the Bonferroni correction method).

Fig.4: Fine-mapped GWAS variants for immune-related diseases are enriched in CORE

a Enrichment analysis of fine-mapped GWAS variants for immune-related diseases or non-immune-related traits in CORE and SE. Left: within CORE, Right: within SE. SLE: Systemic Lupus Erythematosus, CLL: Chronic Lymphocytic Leukemia, T1D: Type 1 Diabetes, Asthma: Asthma, MS: Multiple Sclerosis, Amnesia: Amnesia, DED: Dry Eye Disease, CRC: Colorectal Cancer, Menar.: Menarche, FAI: Fasting Insulin, OC: Ovarian Cancer, Alcoholism: Alcoholism. **b** Enrichment analysis of Young Lab-curated (Hnisz *et al.*) and fine-mapped GWAS variants for Asthma in CORE and SE. Lines are colored by region categories: CORE (red color), SE (blue color), and TE (grey color). B: B cells, CD14: CD14+ Monocytes, CD4: CD4+ Memory T cells, CD8: CD8+ Effector T cells. **c** Enrichment analysis of Young Lab-curated (Hnisz *et al.*) and fine-mapped GWAS variants for CLL in CORE and SE. Lines are colored by region categories: CORE (red color), SE (blue color), and TE (grey color). B: B cells, CD14: CD14+ Monocytes, CD4: CD4+ Memory T cells, CD8: CD8+ Effector T cells. **d** Proportion of differentially expressed genes between CLL and Normal B cells overlapping CORE- or SE-associated genes. UP: up-regulated genes in CLL, DO: down-regulated genes in CLL. CORE_potential: CORE from the potential option (red color), CORE_active: CORE from the active option (orange color), SE: SE (blue color). **e** Box plot showing a comparison of the *RXRA* gene expression between CLL and Normal B cells. Read counts were normalized by the Relative Log Expression (RLE) method. P-value was calculated using the Wald test in DESeq2.

Fig.5: enCORE highlights lineage-committing regulatory determinants in bone marrow hematopoiesis

a Graph-based representation of the similarity among cell type-specific CORE sets identified in BMMC. Left: CORE from the active option. Right: CORE from the potential option. Nodes are colored and annotated by cell types. Edges are shown only for node pairs with cosine similarity greater than 0.55. HSC: Hematopoietic Stem Cells, GMP: Granulocyte-Myeloid Progenitors, CLP: Common Lymphoid Progenitors, MEP: Megakaryocyte-Erythroid Progenitors, CD14_Mono: CD14+ Monocytes, pDC: plasmacytoid Dendritic cells, Pro_B: Pro-B cells, Pre_B: Pre-B cells, Imm_B: Immature B cells, Erythroid: Erythroid cells, NK: NK cells, CD8_Teff: CD8+ Effector T cells, CD8_Tnai: CD8+ naïve T cells, CD4_Tmem: CD4+ Memory T cells, CD4_Tnai: CD4+ naïve T cells. **b** Heatmap for lineage-committing master regulators. Black: gene present in the CORE-associated genes of the given cell type. White:

absent. **c** UMAP coordinates with stream plot of velocity vectors calculated by MultiVelo in CD34+ hematopoietic cells, colored and annotated by cell types. LMPP: Lymphoid-Myeloid Primed Progenitors, HSC: Hematopoietic Stem Cells, MEP: Megakaryocyte-Erythroid Progenitors, MPP: Multipotent Progenitors, Erythrocytes: Erythrocytes, GMP: Granulocyte-Myeloid Progenitors, Prog MK: Megakaryocyte Progenitors, Granulocyte: Granulocytes, Prog DC: Dendritic cell Progenitors, Prog B: B cell Progenitors, Platelet: Platelets. **d** Model of bone marrow hematopoiesis. HSC: Hematopoietic Stem Cells, MPP: Multipotent Progenitors, LMPP: Lymphoid-Myeloid Primed Progenitors, GMP: Granulocyte-Myeloid Progenitors, CLP: Common Lymphoid Progenitors, ETP: Early T-cell Precursors, MEP: Megakaryocyte-Erythroid Progenitors, Mono: Monocytes, pDC: plasmacytoid Dendritic cells, B: B cells, T: T cells, Ery: Erythroid cells.

Fig.6: enCORE identifies CRC-associated epigenetic reprogramming and candidate therapeutic targets.

a NMF analysis of H3K27ac ChIP-seq signals within CORE to distinguish between CRC and Normal samples. Dots are colored by disease state: CRC (red color), Normal (blue color). **b** Log2-transformed fold changes of H3K27ac ChIP-seq signals between CRC and Normal within CORE, SE, and TE. Dots and borderlines are colored by region categories: CORE (red color), SE (blue color), TE (grey color). **c** Enrichment analysis of *de novo* TF motifs in CORE and non-CORE constituents. Left: TF motif enrichment analysis using non-CORE constituents, Right: TF motif enrichment analysis using CORE constituents. The *FOXM1* motif is highlighted in red color. **d** TCGA survival analysis for *MACC1* and *EGFR* genes. P-values were calculated using the Log-rank test. Lines are colored by expression level status: High gene expression (red color), Low gene expression (blue color). Survival curves for the *MACC1* and *EGFR* genes were obtained from RNA-seq and RPPA data, respectively. **e** Track visualization at the *USP7* locus with scATAC-seq signals normalized by reads in TSS. Colored and annotated by disease state: CRC (Colorectal cancer epithelial cells; red color), Normal (Normal epithelial cells; blue color).

Extended Data Fig.1: Cell type-specific transcriptional regulation captured by potential-option

CORE and SE

a Chow-Ruskey diagram of Non-CORE Active Enhancer (non-CORE)-associated genes and Clustered Open Regulatory Element (CORE)-associated genes in PBMC. B cells (red border), Monocytes (purple border), T cells (green border), NK cells (brown border). Monocytes represent the union of CORE-associated genes derived from CD14+ Monocytes and CD16+ Monocytes, while T cells represent the combined set of CORE-associated genes derived from CD4 Memory T cells and CD8 Effector T cells. The color of the borders around each intersection corresponds to the cell types whose genes overlap. The area of each intersection is proportional to the number of genes within the intersection. **b** Chow-Ruskey diagram of typical enhancer (TE)-associated genes and super-enhancer (SE)-associated genes. B cells (red border), Monocytes (purple border), T cells (green border), NK cells (brown border). Monocytes represent the union of SE-associated genes derived from CD14+ Monocytes and CD16+ Monocytes, while T cells represent the combined set of SE-associated genes derived from CD4+ T cells and CD8+ T cells. The color of the borders around each intersection corresponds to the cell types whose genes overlap. The area of each intersection is proportional to the number of genes within the intersection. **c** Log2-transformed fold changes of normalized TPM of genes in each cell type. Normalized TPM values were retrieved from the PBMC Monaco dataset in the Human Protein Atlas (HPA). Comparison between non-CORE-associated genes and CORE-associated genes using the potential option. P-value was calculated by a two-sided Wilcoxon signed-rank test. **d** Dot plot showing the classification performance of CORE in distinguishing SE and TE. Dots are colored by Area Under Curve (AUC), and the dot sizes indicate the specificity. SE derived from B: B cells, CD14: CD14+ Monocytes, CD16: CD16+ Monocytes, CD4: CD4+ T cells, CD4M: CD4+ Memory T cells, CD8: CD8+ T cells, NK: NK cells. **e** Representative receiver operating characteristic (ROC) curves from CD16+ Monocytes. Left: CORE from the potential option, Right: CORE from the active option. The AUC values are indicated in the lower right corner of each plot.

Extended Data Fig.2: Identification of SE from PBMC H3K27ac scCUT&Tag-seq data

a UMAP embedding of the PBMC scCUT&Tag-seq dataset, colored and annotated by clusters. **b** UMAP embedding of the PBMC scCUT&Tag-seq dataset, colored and annotated by cell types. CD14_Mono:

CD14+ Monocytes, CD4_Tmem: CD4+ Memory T cells, CD8_Teff: CD8+ Effector T cells, B: B cells, NK: NK cells. **c** Chow-Ruskey diagram of typical enhancer (TE)-associated genes and super-enhancer (SE)-associated genes. B cells (red border), Monocytes (purple border), T cells (green border), NK cells (brown border). Monocytes represent the union of SE-associated genes derived from CD14+ Monocytes, while T cells represent the combined set of SE-associated genes derived from CD4+ Memory T cells and CD8+ Effector T cells. The color of the borders around each intersection corresponds to the cell types whose genes overlap. The area of each intersection is proportional to the number of genes within the intersection. **d** Heatmap for cell type-specific master regulators. Black: gene present in the SE-associated genes of the given cell type. White: absent. **e** Log2-transformed fold changes of normalized TPM of genes in each cell type. Normalized TPM values were retrieved from the PBMC Monaco dataset in the Human Protein Atlas (HPA). Comparison between TE-associated genes and SE-associated genes. P-value was calculated by a two-sided Wilcoxon signed-rank test.

Extended Data Fig.3: Enrichment of chromatin interactions within CORE in additional cell types

a Log10-transformed weighted degrees of constituents within CORE and non-CORE. Data was retrieved from the H3K27ac HiChIP dataset for Th17 cells. Left: CORE from the active option, Right: CORE from the potential option. **b** Log10-transformed weighted degrees of constituents within CORE and non-CORE. Data was retrieved from the H3K27ac HiChIP dataset for THP-1 cells. Left: CORE from the active option, Right: CORE from the potential option. **c** Log10-transformed weighted degrees of H3K4me1 peaks in each enhancer category from CD14+ Monocytes. Left: active enhancers, Middle: poised enhancers, Right: primed enhancers. Peak group marked by NON represents H3K4me1 peaks that do not fall into either CORE or SE. Data was retrieved from the Hi-C dataset for CD14+ Monocytes. Only Hi-C interactions overlapping with H3K4me1 peaks are considered. CORE from the potential option. P-values were calculated to measure statistical significance (**a,b**, two-sided Wilcoxon rank sum test; **c**, Kruskal-Wallis test with Dunn's test; P-values were adjusted by the Bonferroni correction method).

Extended Data Fig.4: Fine-mapped variants for immune-related diseases are enriched in CORE

across GWAS cohorts

a Enrichment analysis of fine-mapped GWAS variants for immune-related diseases or non-immune-related traits in CORE and SE. Left: within CORE, Right: within SE. SLE: Systemic Lupus Erythematosus, CLL: Chronic Lymphocytic Leukemia, T1D: Type 1 Diabetes, Asthma: Asthma, MS: Multiple Sclerosis, Amnesia: Amnesia, DED: Dry Eye Disease, CRC: Colorectal Cancer, Menar.: Menarche, FAI: Fasting Insulin, OC: Ovarian Cancer, Alcoholism: Alcoholism. **b** Enrichment analysis of fine-mapped GWAS variants for SLE in CORE and SE. Lines are colored by region categories: CORE (red color), SE (blue color), and TE (grey color). B: B cells, CD14: CD14+ Monocytes, CD4: CD4+ Memory T cells, CD8: CD8+ Effector T cells. **c** Enrichment analysis of fine-mapped GWAS variants for T1D in CORE and SE. Lines are colored by region categories: CORE (red color), SE (blue color), and TE (grey color). B: B cells, CD14: CD14+ Monocytes, CD4: CD4+ Memory T cells, CD8: CD8+ Effector T cells. **d** Enrichment analysis of fine-mapped GWAS variants for Asthma in CORE and SE. Lines are colored by region categories: CORE (red color), SE (blue color), and TE (grey color). B: B cells, CD14: CD14+ Monocytes, CD4: CD4+ Memory T cells, CD8: CD8+ Effector T cells. **e** Enrichment analysis of fine-mapped GWAS variants for MS in CORE and SE. Lines are colored by region categories: CORE (red color), SE (blue color), and TE (grey color). B: B cells, CD14: CD14+ Monocytes, CD4: CD4+ Memory T cells, CD8: CD8+ Effector T cells.

Extended Data Fig.5: Enrichment of fine-mapped CAD risk loci within CORE

a Enrichment analysis of fine-mapped coronary artery disease (CAD) risk loci from Webb et al. GWAS cohort in CORE and SE. CORE (red color), SE (blue color), and TE (grey color). B: B cells, CD14: CD14+ Monocytes, CD4: CD4+ Memory T cells, CD8: CD8+ Effector T cells. **b** Enrichment analysis of fine-mapped CAD risk loci from Van der Harst et al. GWAS cohort in CORE and SE. CORE (red color), SE (blue color), and TE (grey color). B: B cells, CD14: CD14+ Monocytes, CD4: CD4+ Memory T cells, CD8: CD8+ Effector T cells. **c** Enrichment analysis of fine-mapped multi-variate CAD and TAG risk loci from Siewert et al. GWAS cohort in CORE and SE. CORE (red color), SE (blue color), and TE (grey color). B: B cells, CD14: CD14+ Monocytes, CD4: CD4+ Memory T cells, CD8: CD8+ Effector T cells. **d** Cell type specificity of fine-mapped CAD risk loci enrichment. Borderlines are colored by cell types:

CD8_Teff (CD8+ Effector T cells; green color), CD4_Tmem (CD4+ Memory T cells), B (B cells; red color), CD14_Mono (CD14+ Monocytes; purple color). **e** Fine-mapped CAD risk loci near the *LIPA* gene. The *LIPA* gene is highlighted in red.

Extended Data Fig.6: CORE constituents delineate lineage-specific epigenetic state

a Graph-based representation of the similarity among CORE constituents identified across different cell types in BMMC. Left: CORE constituents from the active option. Right: CORE constituents from the potential option. Nodes are colored and annotated by cell types. Edges are shown only for node pairs with cosine similarity greater than 0.55. HSC: Hematopoietic Stem Cells, GMP: Granulocyte-Myeloid Progenitors, CLP: Common Lymphoid Progenitors, MEP: Megakaryocyte-Erythroid Progenitors, CD14_Mono: CD14+ Monocytes, pDC: plasmacytoid Dendritic cells, Pro_B: Pro-B cells, Pre_B: Pre-B cells, Imm_B: Immature B cells, Erythroid: Erythroid cells, NK: NK cells, CD8_Teff: CD8+ Effector T cells, CD8_Tnai: CD8+ naïve T cells, CD4_Tmem: CD4+ Memory T cells, CD4_Tnai: CD4+ naïve T cells. **b** Graph-based representation of the similarity among non-CORE constituents identified across different cell types in BMMC. Left: non-CORE constituents from the active option. Right: non-CORE constituents from the potential option. Nodes are colored and annotated by cell types. Edges are shown only for node pairs with cosine similarity greater than 0.55. HSC: Hematopoietic Stem Cells, GMP: Granulocyte-Myeloid Progenitors, CLP: Common Lymphoid Progenitors, MEP: Megakaryocyte-Erythroid Progenitors, CD14_Mono: CD14+ Monocytes, pDC: plasmacytoid Dendritic cells, Pro_B: Pro-B cells, Pre_B: Pre-B cells, Imm_B: Immature B cells, Erythroid: Erythroid cells, NK: NK cells, CD8_Teff: CD8+ Effector T cells, CD8_Tnai: CD8+ naïve T cells, CD4_Tmem: CD4+ Memory T cells, CD4_Tnai: CD4+ naïve T cells.

Extended Data Fig.7: enCORE captures hematopoietic lineage-matched BM-to-PB maturation trajectories

a Heatmap for cosine similarity metrics between CORE from bone marrow (BM) and peripheral blood (PB). Red: cosine similarity > 0.55. White: cosine similarity < 0.55. Values in each heatmap cell represent the cosine similarity using the potential option of enCORE. Heatmap cells with dotted borders

indicate cases where the cosine similarity < 0.55 using the active option of enCORE. HSC: Hematopoietic Stem Cells, GMP: Granulocyte-Myeloid Progenitors, CLP: Common Lymphoid Progenitors, CD14_Mono: CD14+ Monocytes, CD16_Mono: CD16+ Monocytes, cDC: classical Dendritic cells, B: B cells, Pro_B: Pro-B cells, Pre_B: Pre-B cells, Imm_B: Immature B cells, NK: NK cells, CD8_Teff: CD8+ Effector T cells, CD8_Tnai: CD8+ naïve T cells, CD4_Tmem: CD4+ Memory T cells, CD4_Tnai: CD4+ naïve T cells. **b** Dot plot for the enrichment of monocyte-elevated or macrophage-elevated H3K27ac peaks within CORE and SE. Values shown in each dot represent the percentage of monocyte-elevated or macrophage-elevated H3K27ac peaks that overlap CORE or SE. Dot size is determined by the common odds ratio. Red/Orange: p-value < 0.05. Grey: p-value > 0.05. Mono: monocyte-elevated H3K27ac peaks. Macro: macrophage-elevated H3K27ac peaks. CORE_pot: CORE from CD14+ Monocytes using the potential option. CORE_act: CORE from CD14+ Monocytes using the active option. SE_K27ac: SE from CD14+ Monocytes. **c-e** Track visualization at the **c** *MEF2A*, **d** *BHLHE41*, and **e** *DCSTAMP* loci with normalized H3K27ac ChIP-seq signals in macrophages and monocytes. Bone marrow: CORE from BM CD14+ Monocytes using the potential option. Peripheral blood: CORE from PB CD14+ Monocytes using the potential option. Macro-specific: macrophage-elevated H3K27ac peaks without TSS exclusion. Colored and annotated by cell types: Macro (Macrophages; red color), Mono (Monocytes; purple color).

Extended Data Fig.8: *In silico* knockout of a CORE-associated enhancer near *USP7*

a Track visualization at the *USP7* locus with H3K27ac HiChIP interactions from HT29 cells and normalized H3K27ac ChIP-seq signals in CRC and Normal samples. Colored and annotated by disease state: CRC (red color), Normal (blue color). **b** Track visualization of RNA-seq from the Left Colon ontology under *in silico* knockout of the CORE-associated enhancer near *USP7*. **c** Magnitude of the predicted effect on RNA-seq reads from the Left Colon ontology under the same perturbation.

Extended Data Fig.9: Metastasis-related genes are enriched among CORE-associated genes and linked to distinct node-level network properties

a Proportion of CORE- and SE-associated genes that overlap with metastasis-related genes, relative

to the total number of metastasis-related genes. Target (Bulk): differentially expressed genes between metastatic tumor and primary CRC in bulk RNA-seq data overlapping with target genes in TF regulons from scRNA-seq data, TF (scRNA): differential TFs between metastatic tumor and primary CRC based on TF regulon activity in scRNA-seq data. **b** Box plots showing weighted degree and weighted transitivity (weighted local clustering coefficients) in CORE only, SE only, and Other peaks. CORE only: H3K4me1 peaks located exclusively within CORE, SE only: H3K4me1 peaks located exclusively within SE, Other peaks: H3K4me1 peaks that do not fall into either CORE or SE. Correlation coefficients are used as edge weights in the co-accessibility network. **c** Log2-transformed fold changes of mean weighted degree and mean clustering coefficient between CORE and SE. DEGR: Weighted degree (W.D), LCLC: Weighted transitivity (W.T). **d** Fold changes in the number of metastasis-related genes overlapping genes annotated from peaks with a high weighted local clustering coefficient, relative to those annotated from peaks with a high weighted degree. Exp: expected fold changes, calculated as the number of annotated genes from peaks with a high weighted clustering coefficient divided by the number of annotated genes from peaks with a high weighted degree, Obs: observed fold changes. Target (Bulk): differentially expressed genes between metastatic tumors and primary CRC in bulk RNA-seq data overlapping target genes in TF regulons inferred from scRNA-seq data, TF (scRNA): differential TFs between metastatic tumors and primary CRC based on TF regulon activity in scRNA-seq data.

Supplementary Fig.1: GO enrichment analysis of CORE in CD14+ and CD16+ monocytes

a GO enrichment analysis in CD14+ Monocytes. Shared: Regions shared between CORE and SE, CORE only: Regions exclusively found in CORE, SE only: Regions exclusively found in SE, TE only: Regions exclusively found in TE. Dots are colored by –log10(binomial FDR-adjusted P-value), and the dot sizes indicate the Gene Set Coverage. **b** GO enrichment analysis in CD16+ Monocytes. Shared: Regions shared between CORE and SE, CORE only: Regions exclusively found in CORE, SE only: Regions exclusively found in SE, TE only: Regions exclusively found in TE. Dots are colored by –log10(binomial FDR-adjusted P-value), and the dot sizes indicate the Gene Set Coverage.

Supplementary Fig.2: Track visualization of KLF4, RBPJ, and LYZ in CD14+ monocytes

a Track visualization at the *KLF4* locus with normalized H3K27ac ChIP-seq signals. **b** Track visualization at the *RBPJ* locus with normalized H3K27ac ChIP-seq signals. **c** Track visualization at the *LYZ* locus with normalized H3K27ac ChIP-seq signals. Colored and annotated by region categories: CORE (red color), SE (blue color).

Supplementary Fig.3: Active-option CORE shows a higher active-to-poised enhancer ratio than potential-option CORE

a Bar plots showing the fold change and the number of active enhancers in CD19+ B cells across different CORE calling options. **b** Bar plots showing the fold change and the number of active enhancers in CD14+ Monocytes across different CORE calling options. Left: fold change in the number of active enhancers relative to the number of poised enhancers, Right: The number of active enhancers. Borderlines are colored by CORE calling options: Potential (CORE from the potential option; purple color), 5kb (CORE from the active option using 5kb as proximal enhancer distance; green color), 7.5kb (CORE from the active option using 7.5kb as proximal enhancer distance; red color), 10kb (CORE from the active option using 10kb as proximal enhancer distance; blue color), 25kb (CORE from the active option using 25kb as proximal enhancer distance; orange color).

Supplementary Fig.4: Immune-related disease GWAS variants are enriched in CORE (Hnisz *et al.*)

a Enrichment analysis of GWAS variants for immune-related diseases or non-immune-related traits in CORE and SE. Left: within CORE, Right: within SE. SLE: Systemic Lupus Erythematosus, CLL: Chronic Lymphocytic Leukemia, T1D: Type 1 Diabetes, Asthma: Asthma, MS: Multiple Sclerosis, ALZ: Alzheimer's disease, CRC: Colorectal Cancer, Menar.: Menarche, FAI: Fasting Insulin, OC: Ovarian Cancer, Alcoholism: Alcoholism. **b** Enrichment analysis of GWAS variants for SLE in CORE and SE. CORE (red color), SE (blue color), and TE (grey color). B: B cells, CD14: CD14+ Monocytes, CD4: CD4+ Memory T cells, CD8: CD8+ Effector T cells. **c** Enrichment analysis of GWAS variants for T1D in

CORE and SE. CORE (red color), SE (blue color), and TE (grey color). B: B cells, CD14: CD14+ Monocytes, CD4: CD4+ Memory T cells, CD8: CD8+ Effector T cells. **d** Enrichment analysis of GWAS variants for MS in CORE and SE. CORE (red color), SE (blue color), and TE (grey color). B: B cells, CD14: CD14+ Monocytes, CD4: CD4+ Memory T cells, CD8: CD8+ Effector T cells. **e** Proportion of GWAS variants that overlap with CORE or SE, relative to the total number of GWAS variants in each autoimmune disease category. Borderlines are colored by method: CORE (red color), SE (blue color).

Supplementary Fig.5: UMAP visualization and canonical marker expression in CRC epithelial cells

a UMAP embedding of the CRC scATAC-seq dataset colored by *EPCAM* gene expression. **b** UMAP embedding of the CRC scATAC-seq dataset colored by sub-clusters in epithelial cells. **c** UMAP embedding of the CRC scATAC-seq dataset colored by condition (disease state) information. CRC (Colorectal cancer epithelial cells; red color), Normal (Normal epithelial cells; blue color).

Supplementary Fig.6: Comparison of classification performance between CORE and SE (GSE156613)

a NMF analysis of H3K27ac ChIP-seq signals within SE to distinguish between CRC and Normal samples. Dots are colored by disease state: CRC (red color), Normal (blue color). **b** Violin plot showing classification performance between CORE and SE. Area Under Curve (AUC) values were calculated by K-means clustering on the NMF-reduced dimension. **c-f** Violin plot showing XGBoost classification performance between CORE and SE. **c** AUC (Top Left), **d** F1 score (Top Right), **e** Precision (Bottom Left), and **f** Recall (Bottom Right). P-values were calculated to measure statistical significance (**b-f**, two-sided Wilcoxon rank sum test).

Supplementary Fig.7: Comparison of classification performance between CORE and SE (GSE136889)

a NMF analysis of H3K27ac ChIP-seq signals within SE to distinguish between CRC and Normal

samples. Dots are colored by disease state: CRC (red color), Normal (blue color). **b** Violin plot showing classification performance between CORE and SE. Area Under Curve (AUC) values were calculated by K-means clustering on the NMF-reduced dimension. **c-f** Violin plot showing XGBoost classification performance between CORE and SE. **c** AUC (Top Right), **d** F1 score (Bottom Left), **e** Precision (Bottom Middle), and **f** Recall (Bottom Right). **g** Hierarchical clustering result for CRC and Normal samples. Euclidean distance was used as a distance metric for clustering. The cluster containing the most similar CRC samples identified using SE was labeled as Clustered CRC. **h** Log2-transformed fold changes of H3K27ac ChIP-seq signals between CRC and Normal within CORE, SE, and TE. Dots and borderlines are colored by region categories: CORE (red color), SE (orange color), TE (grey color). P-values were calculated to measure statistical significance (**b-f**, two-sided Wilcoxon rank sum test).

Supplementary Fig.8: CRC master TF candidates are enriched among CORE-associated genes

a Proportion of CRC master TF candidates that overlap with genes associated with each region category, relative to the total number of CRC master TF candidates in each study. Region category: CORE (CORE; red color), SE from GSE136889 (SE_test; orange color), and SE from GSE156613 (SE; blue color). **b** Heatmap of CRC master TF candidates. Black: gene present in the region category-associated genes of the given cell type. White: absent.

Supplementary Fig.9: Survival analysis for CORE-associated genes

a-h Kaplan-Meier survival analysis of CORE-associated genes in TCGA CRC samples. **a** *ANO1*, **b** *ARL4C*, **c** *FOXD1*, **d** *PRDM15*, **e** *HOXC10*, **f** *ISM1*, **g** *BRD2*, **h** *KDM4B*. P-values were calculated using the Log-rank test. Lines are colored by expression level status: High gene expression (red color), Low gene expression (blue color). Survival curves were obtained from RNA-seq data.

Figure 1. Overview of enCORE and its application to single-cell chromatin accessibility datasets

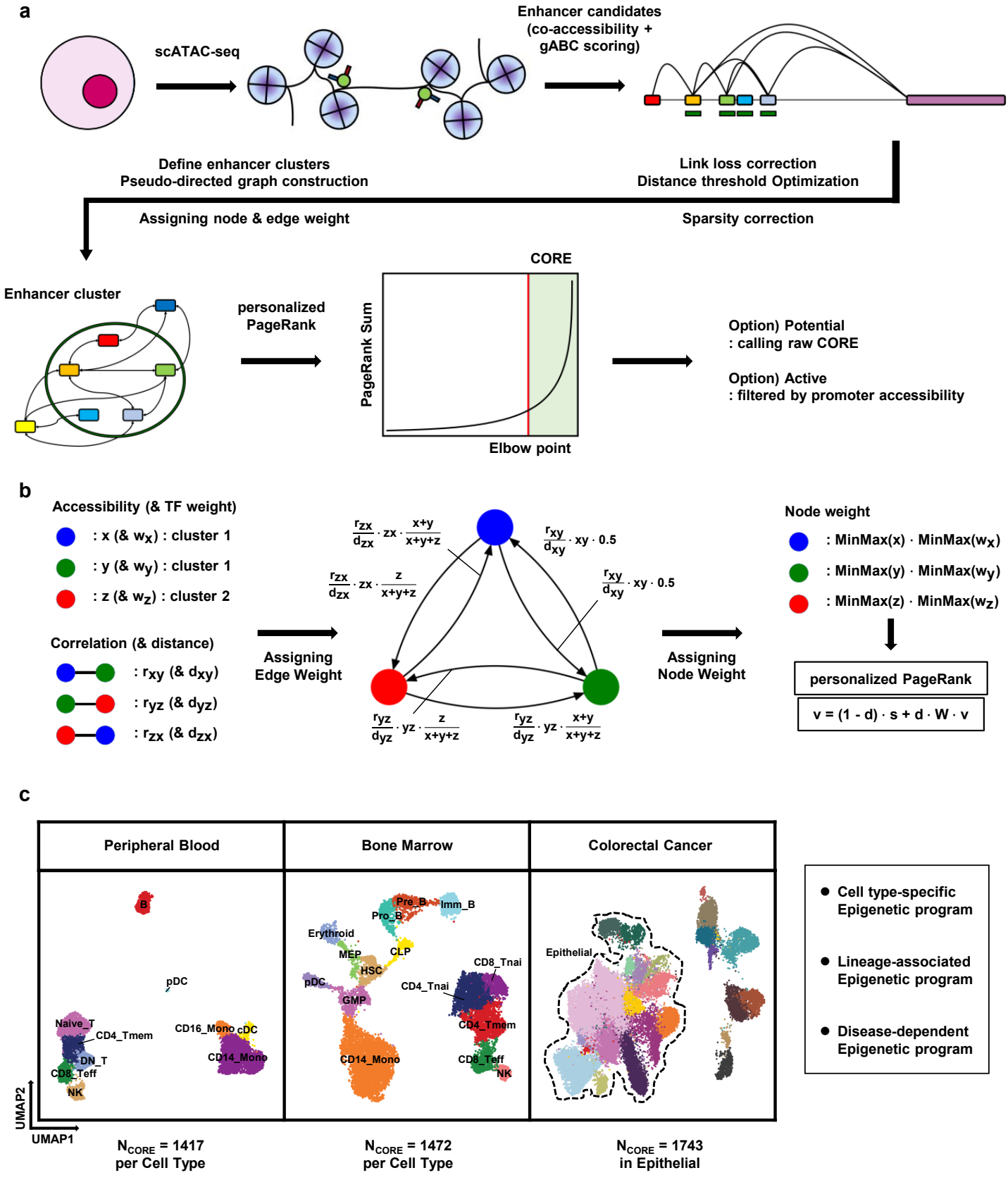


Figure 2. enCORE recapitulates cell type-specific transcriptional regulation in PBMC

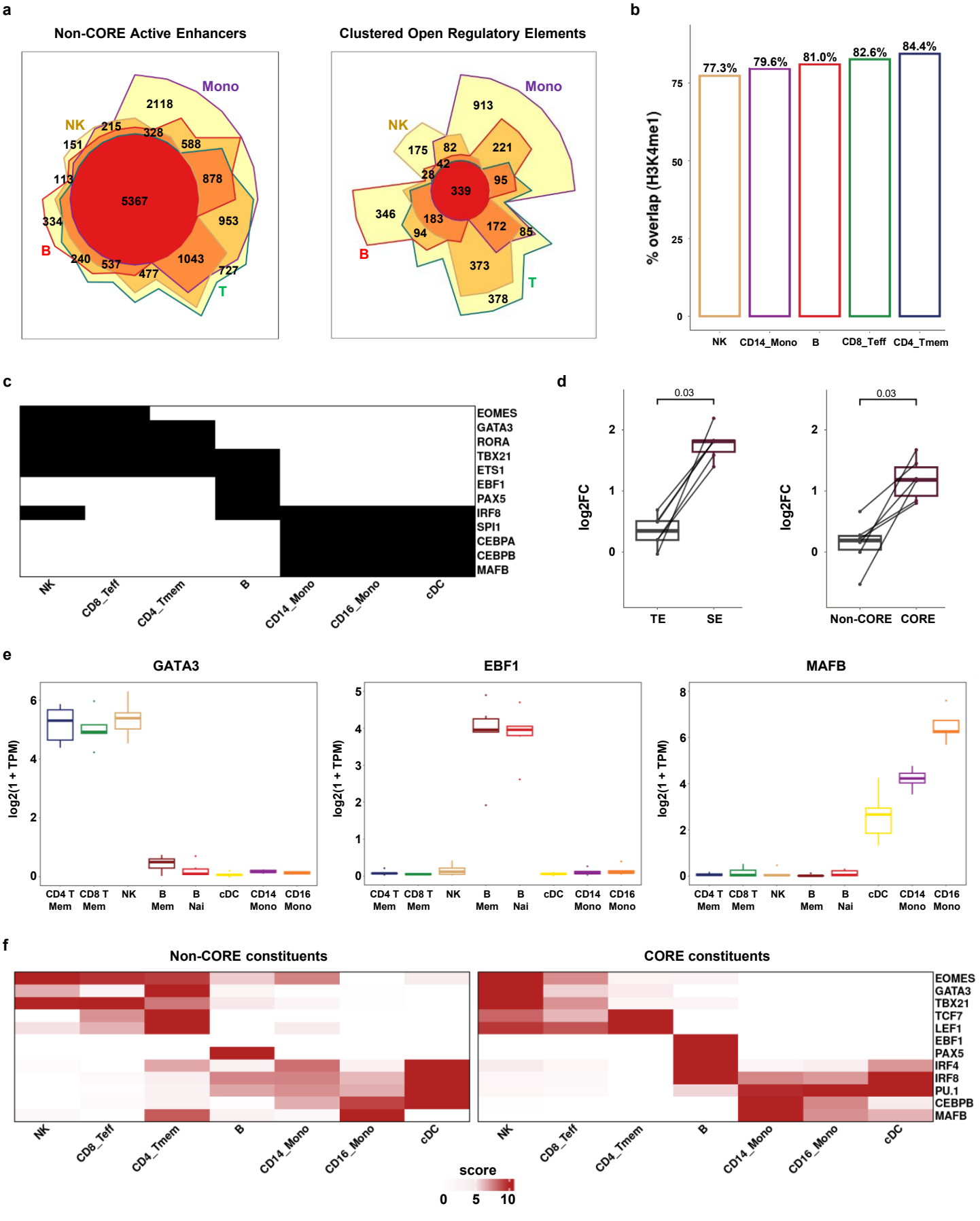


Figure 3. CORE constituents show enriched chromatin interactions

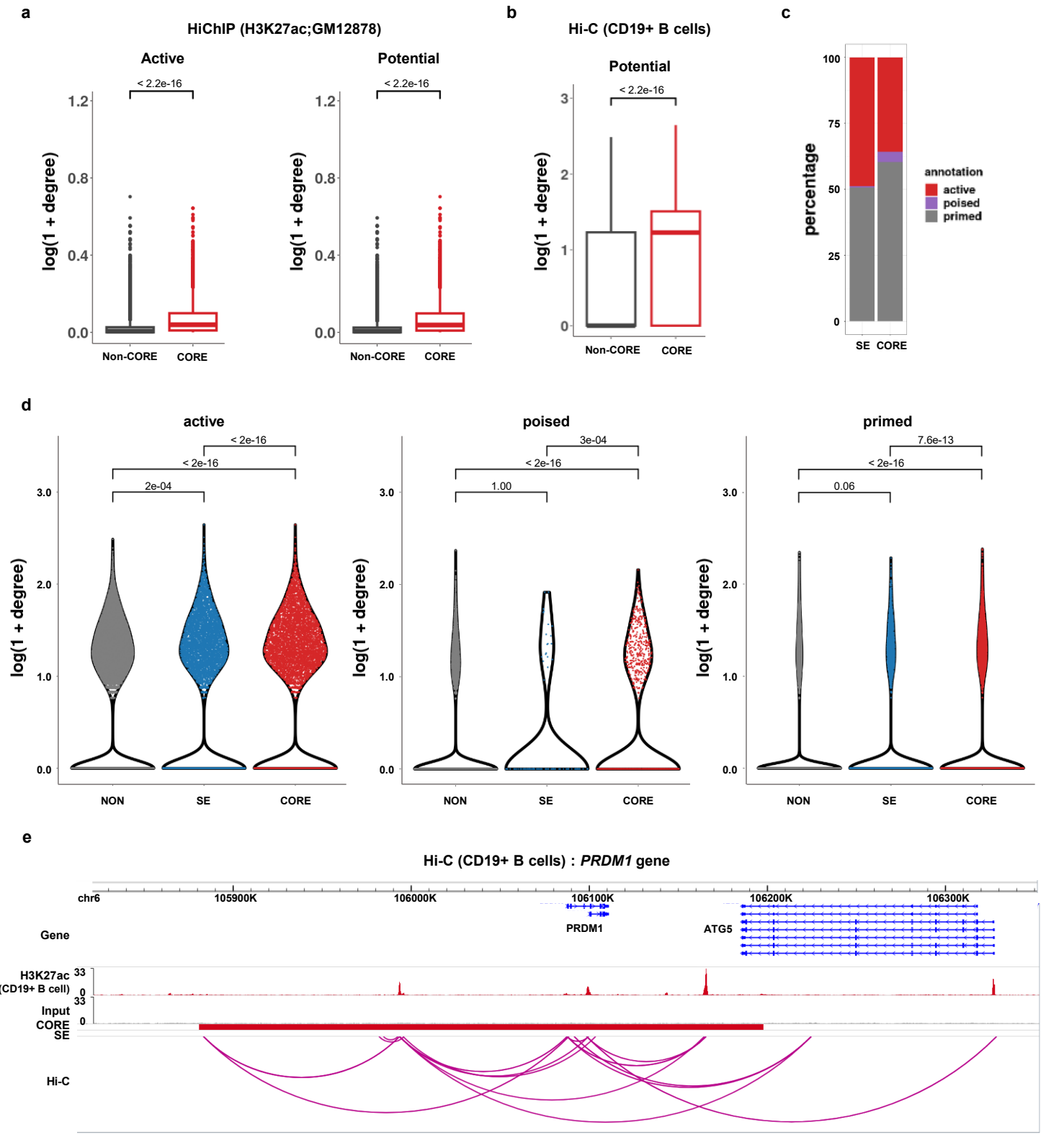


Figure 4. Fine-mapped GWAS variants for immune-related diseases are enriched in CORE

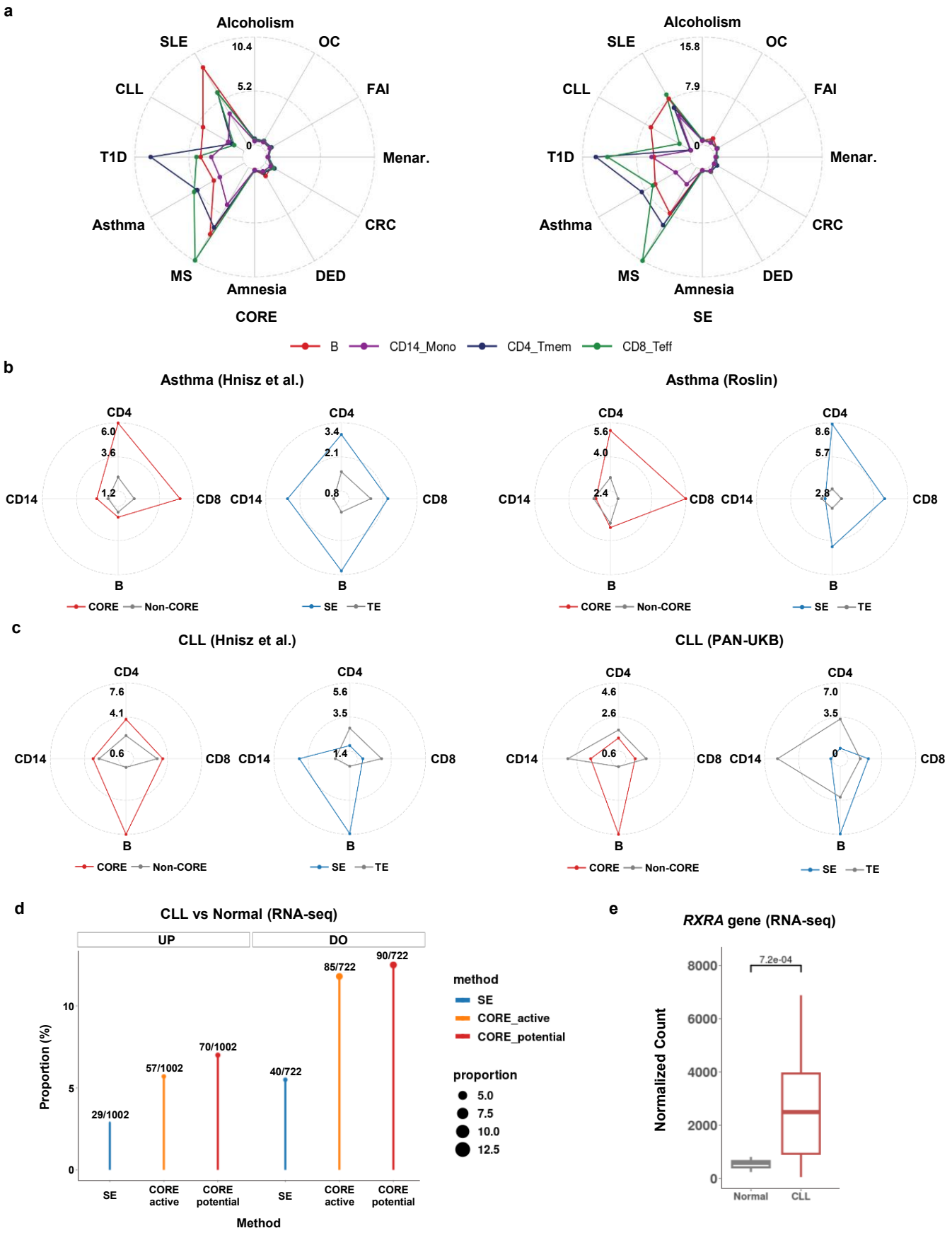


Figure 5. enCORE highlights lineage-committing regulatory determinants in bone marrow hematopoiesis

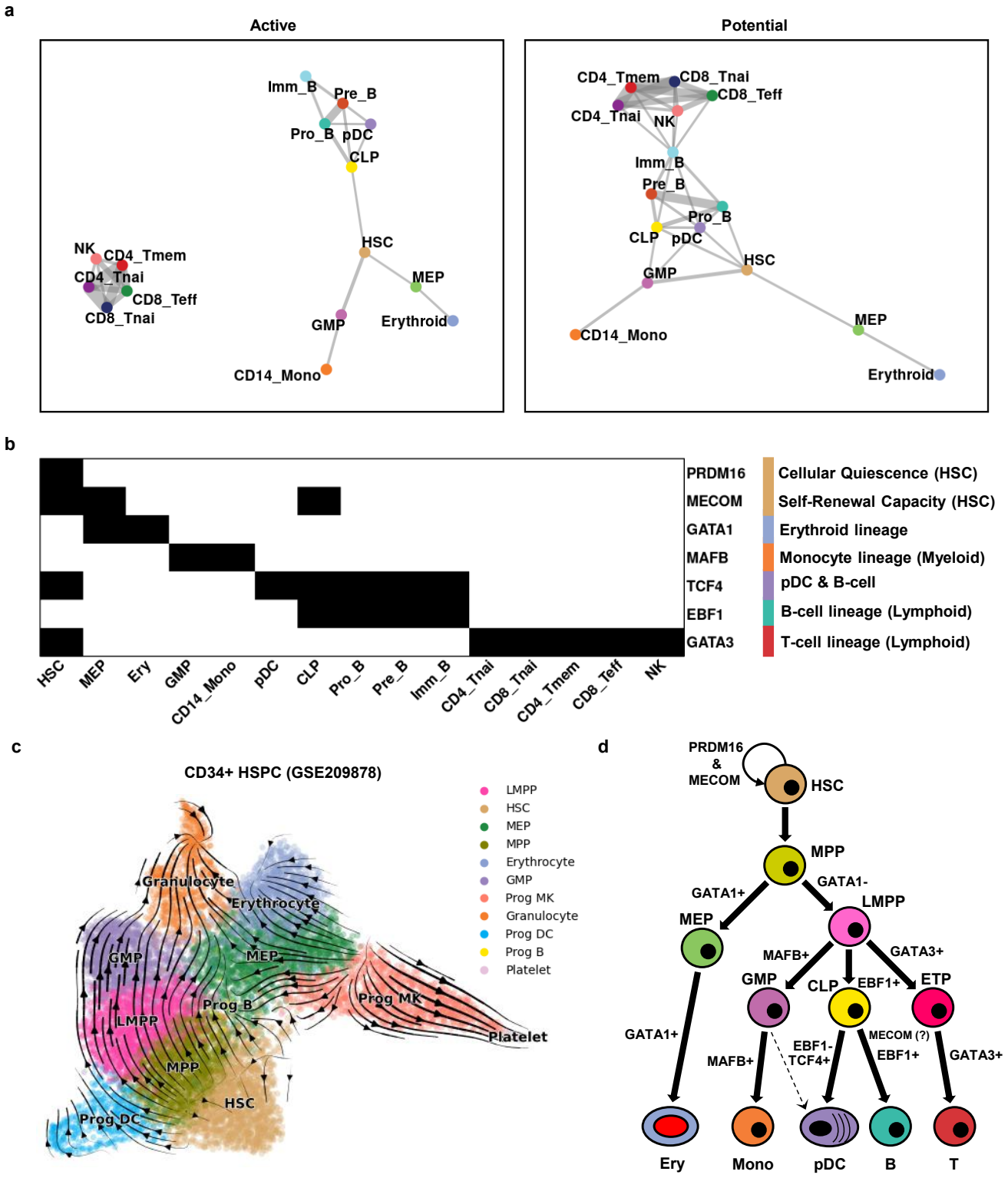
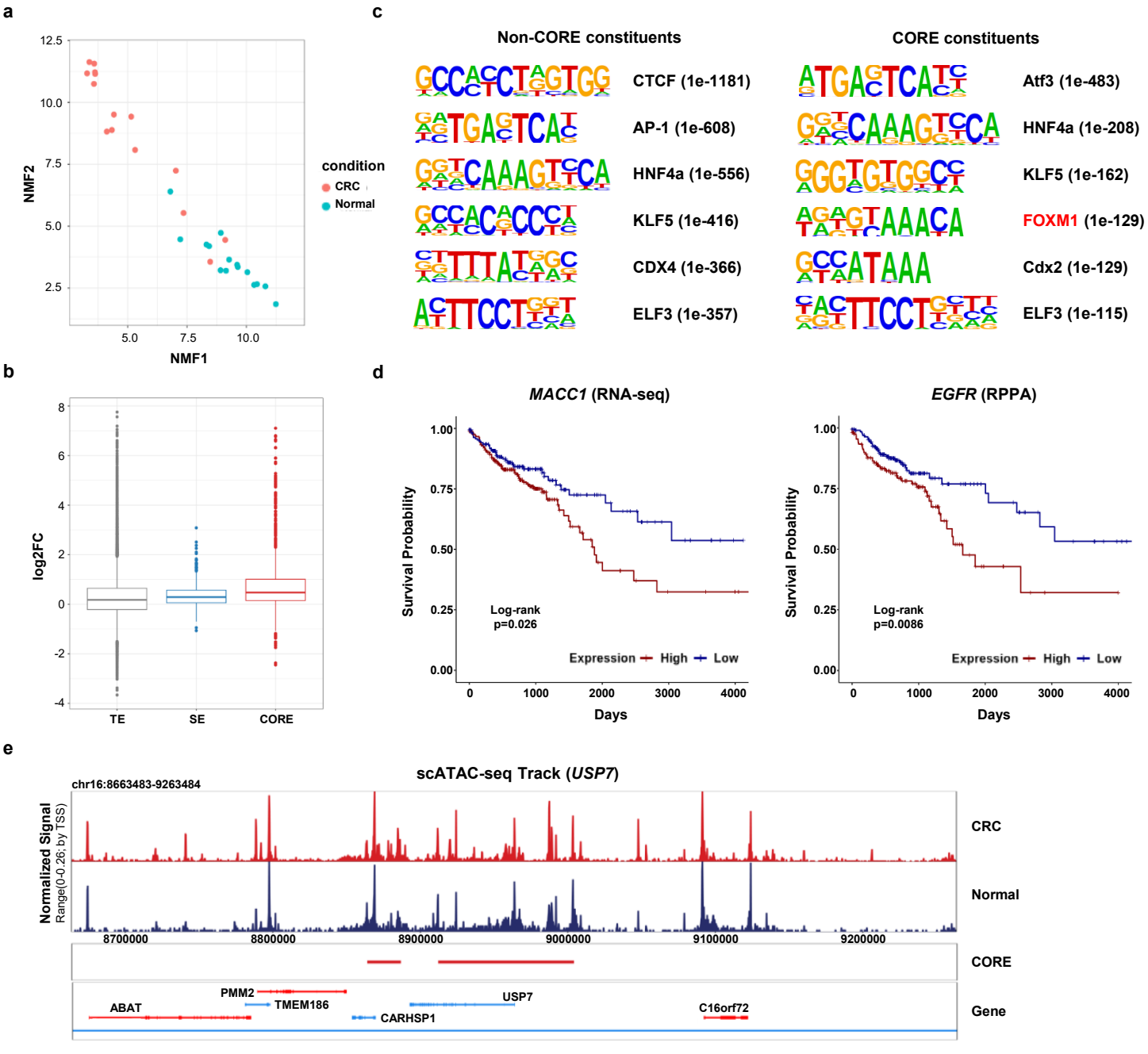
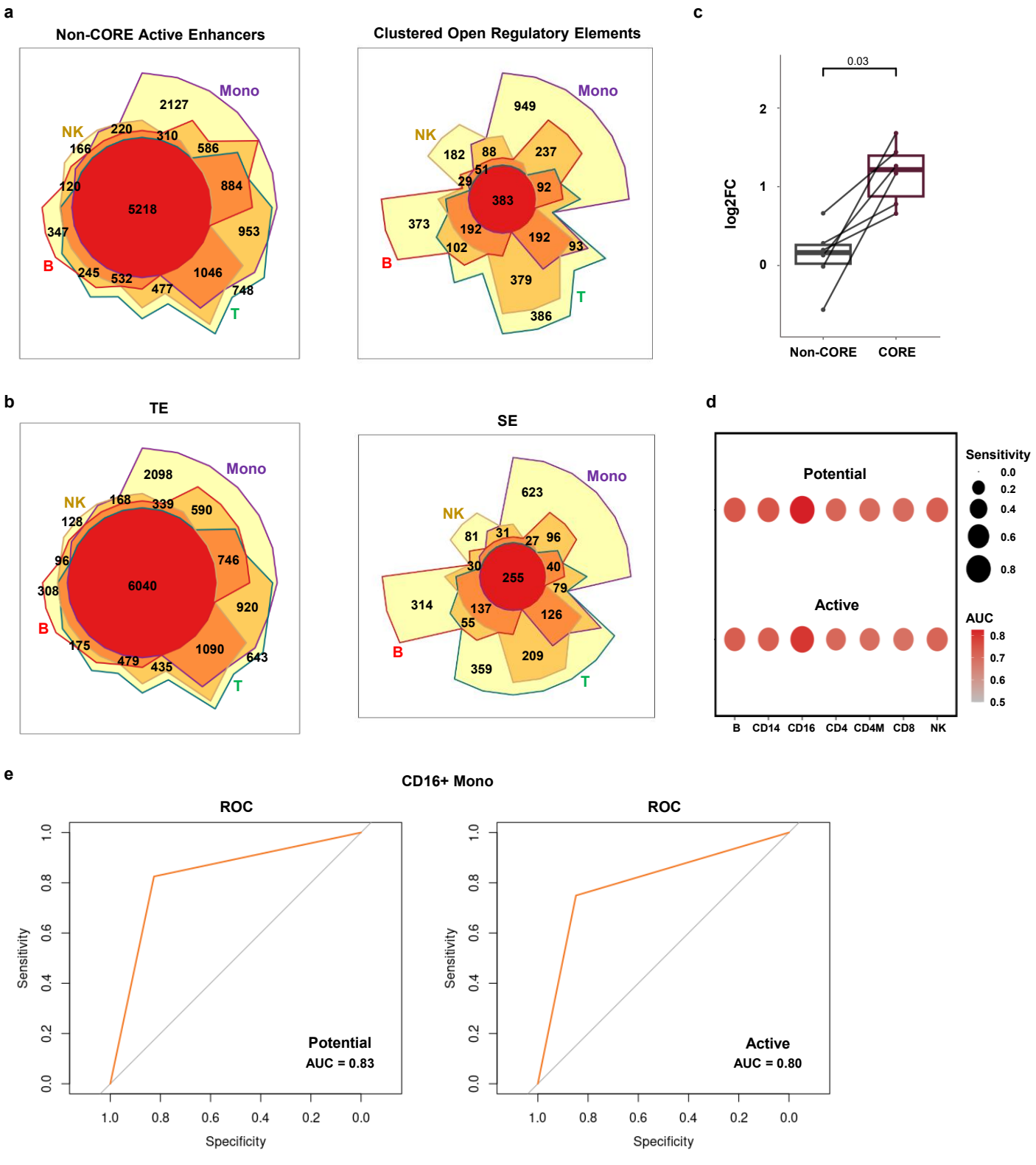


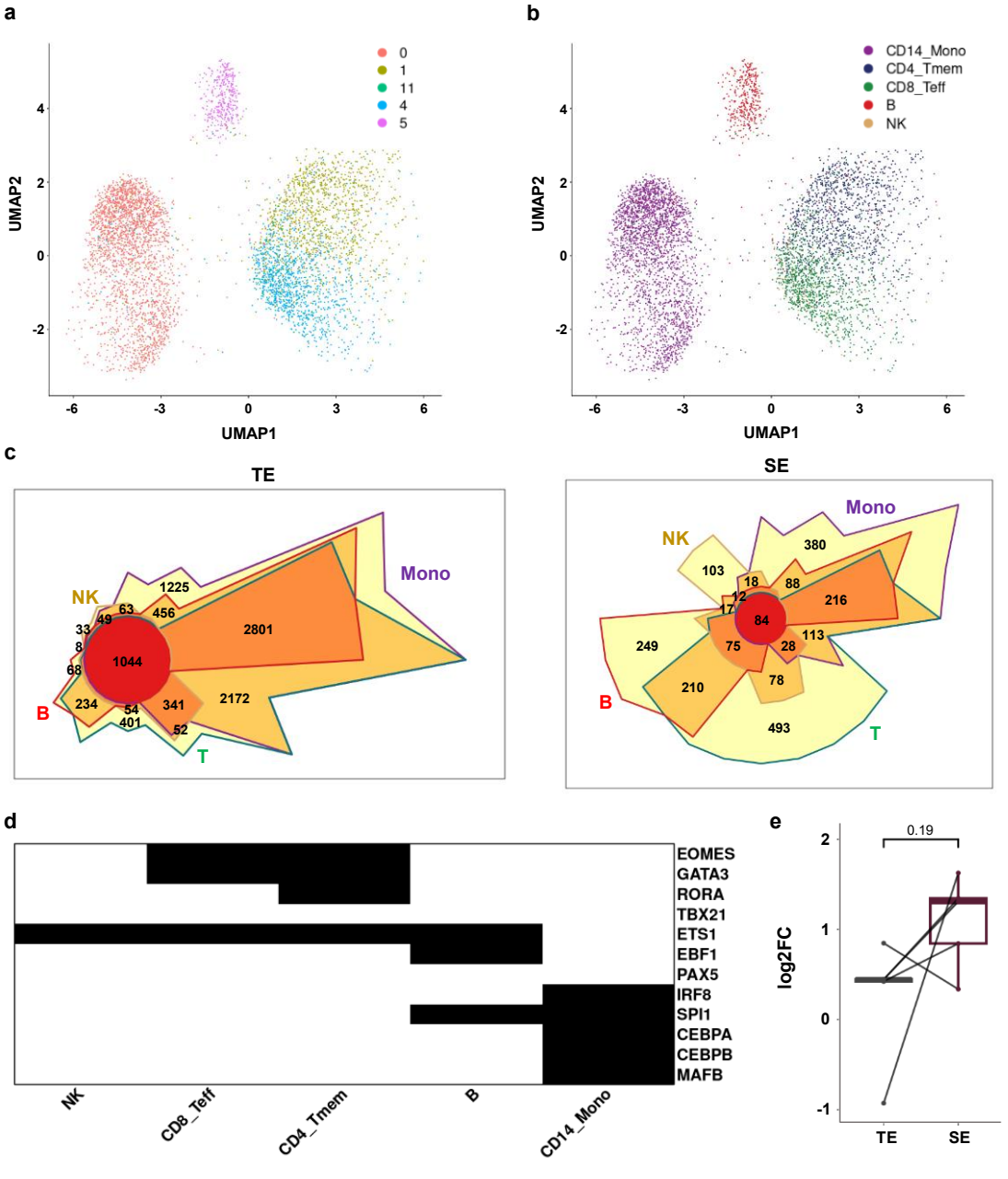
Figure 6. enCORE identifies CRC-associated epigenetic reprogramming and candidate therapeutic targets



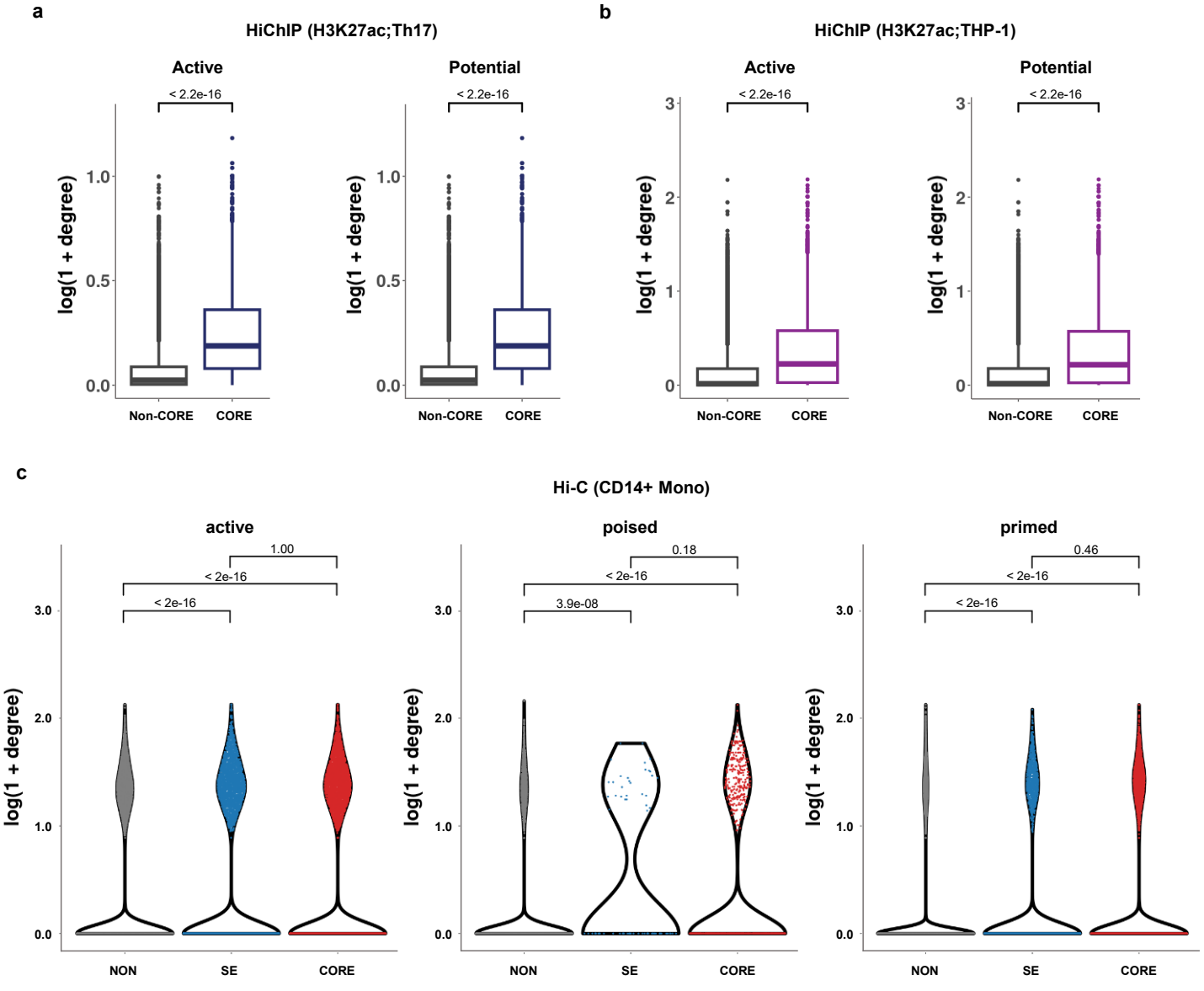
Extended Data Figure 1. Cell type-specific transcriptional regulation captured by potential-option CORE and SE



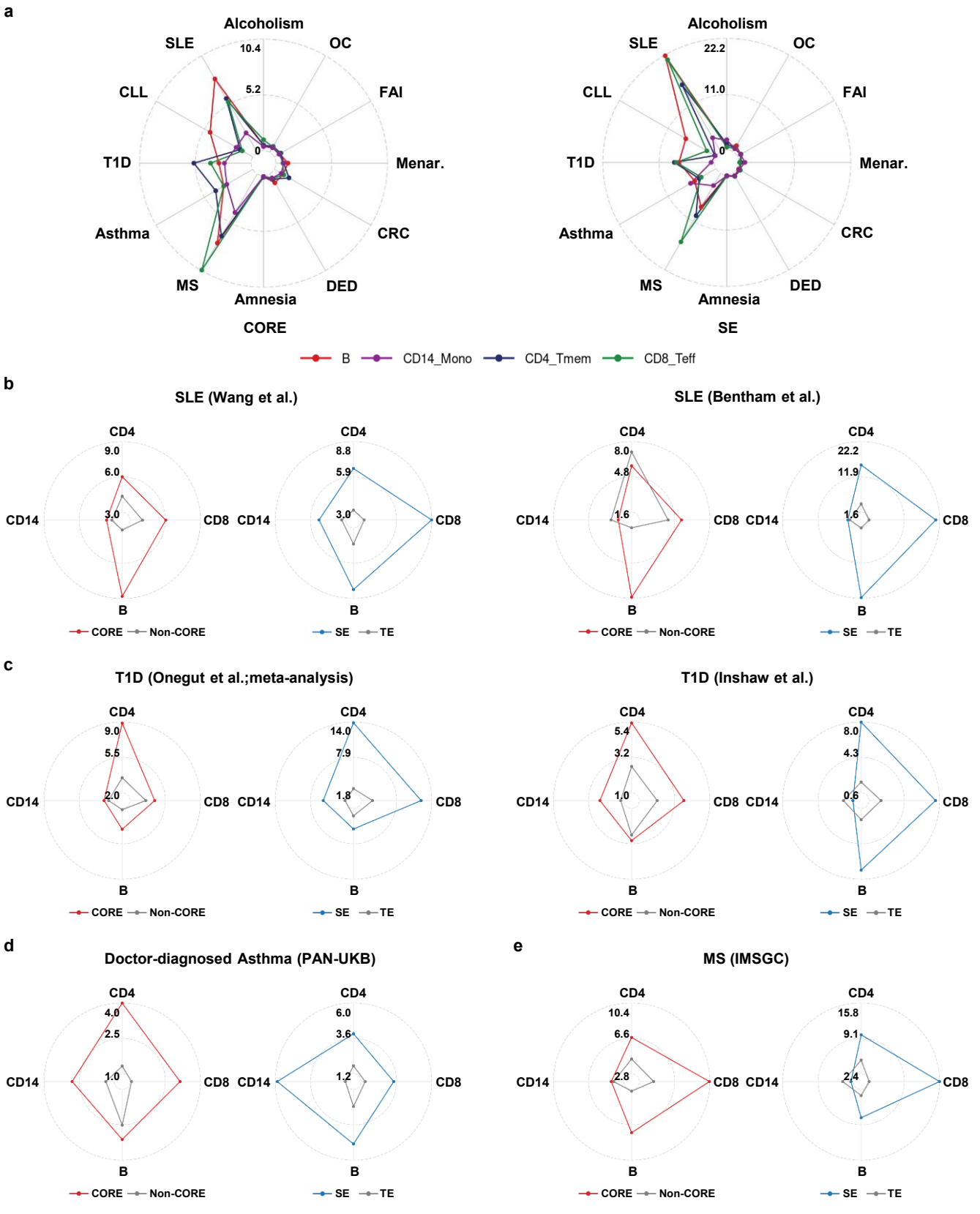
Extended Data Figure 2. Identification of SE from PBMC H3K27ac scCUT&Tag-seq data



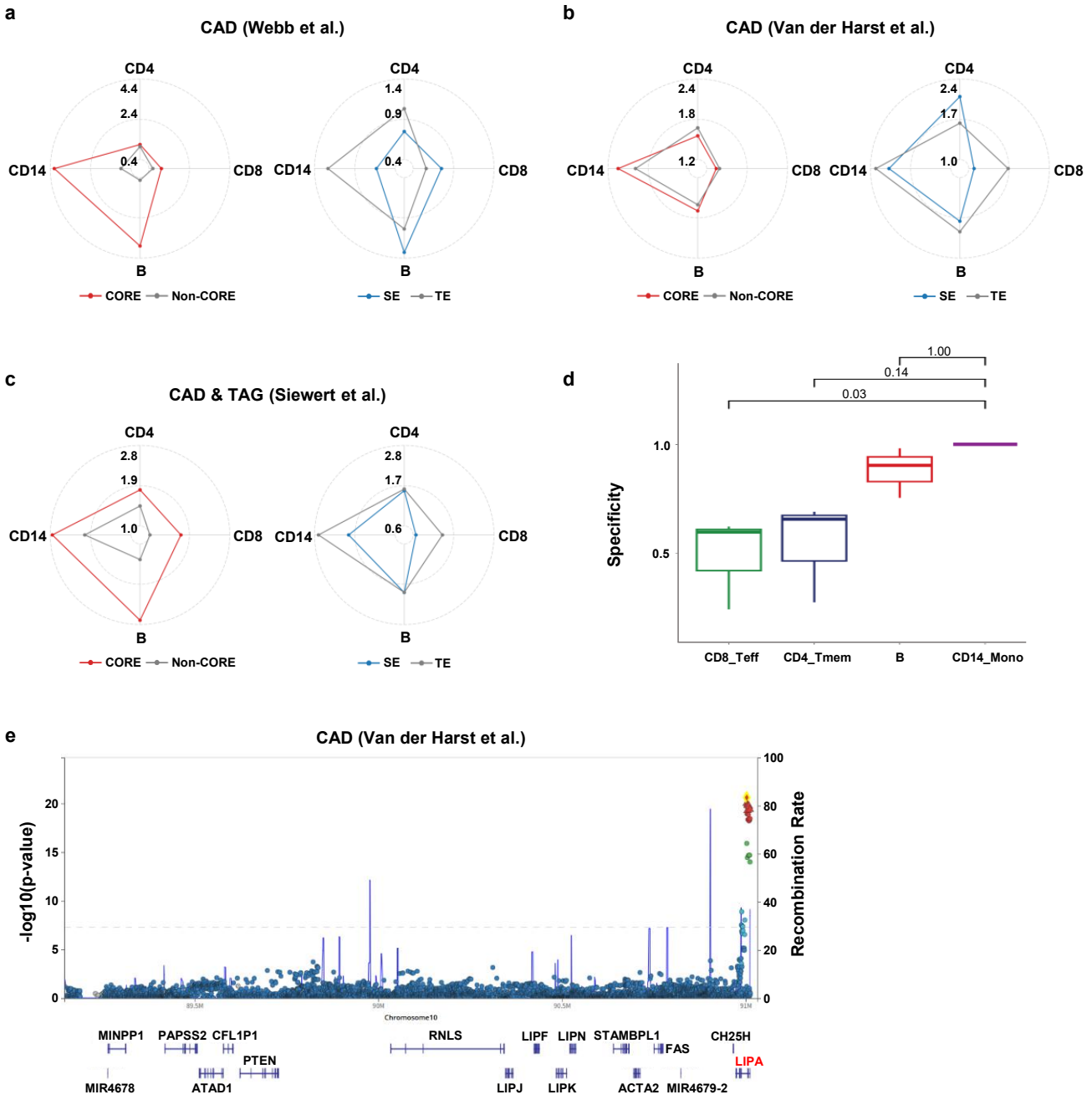
Extended Data Figure 3. Enrichment of chromatin interactions within CORE in additional cell types



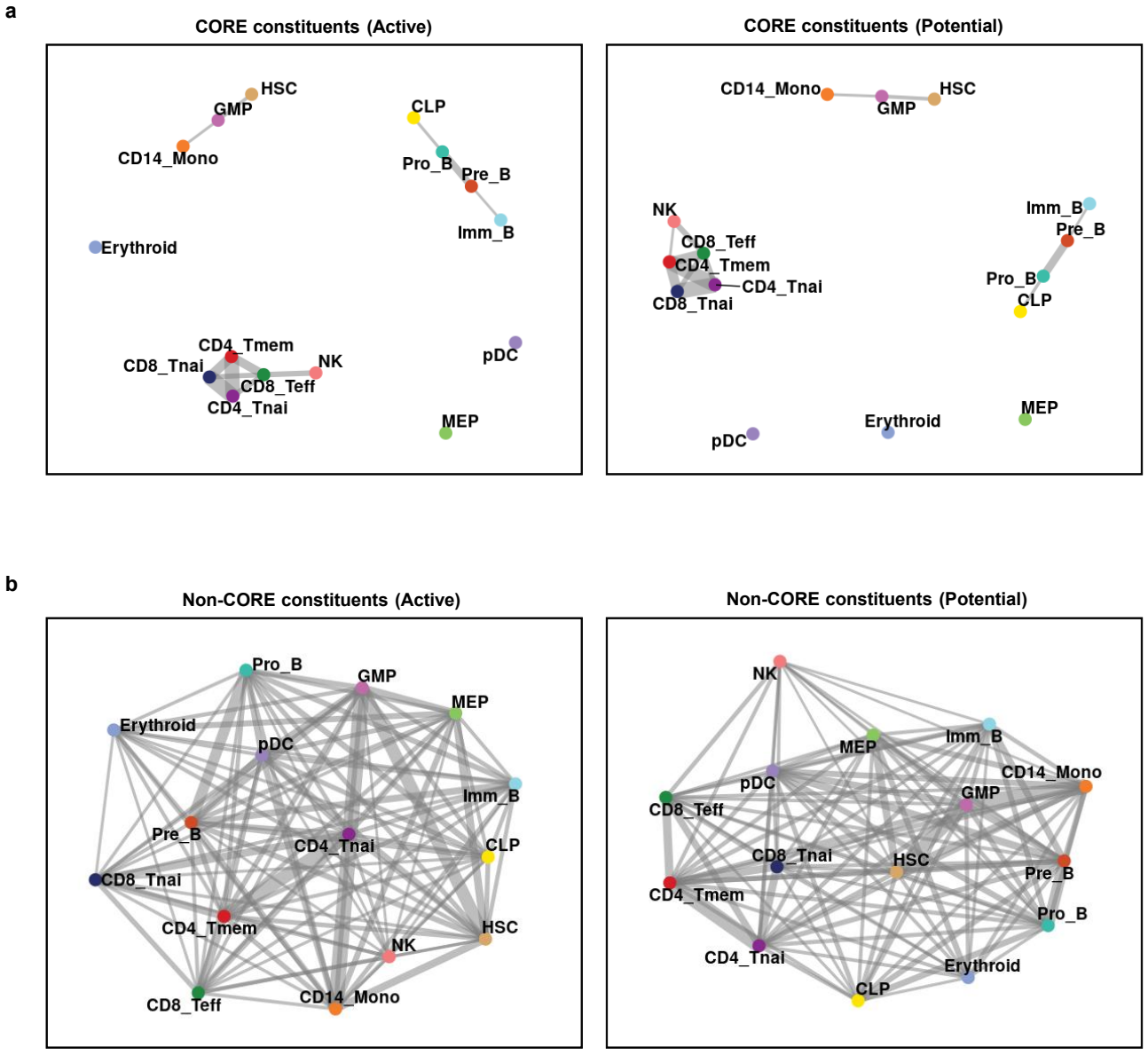
Extended Data Figure 4. Fine-mapped variants for immune-related diseases are enriched in CORE across GWAS cohorts



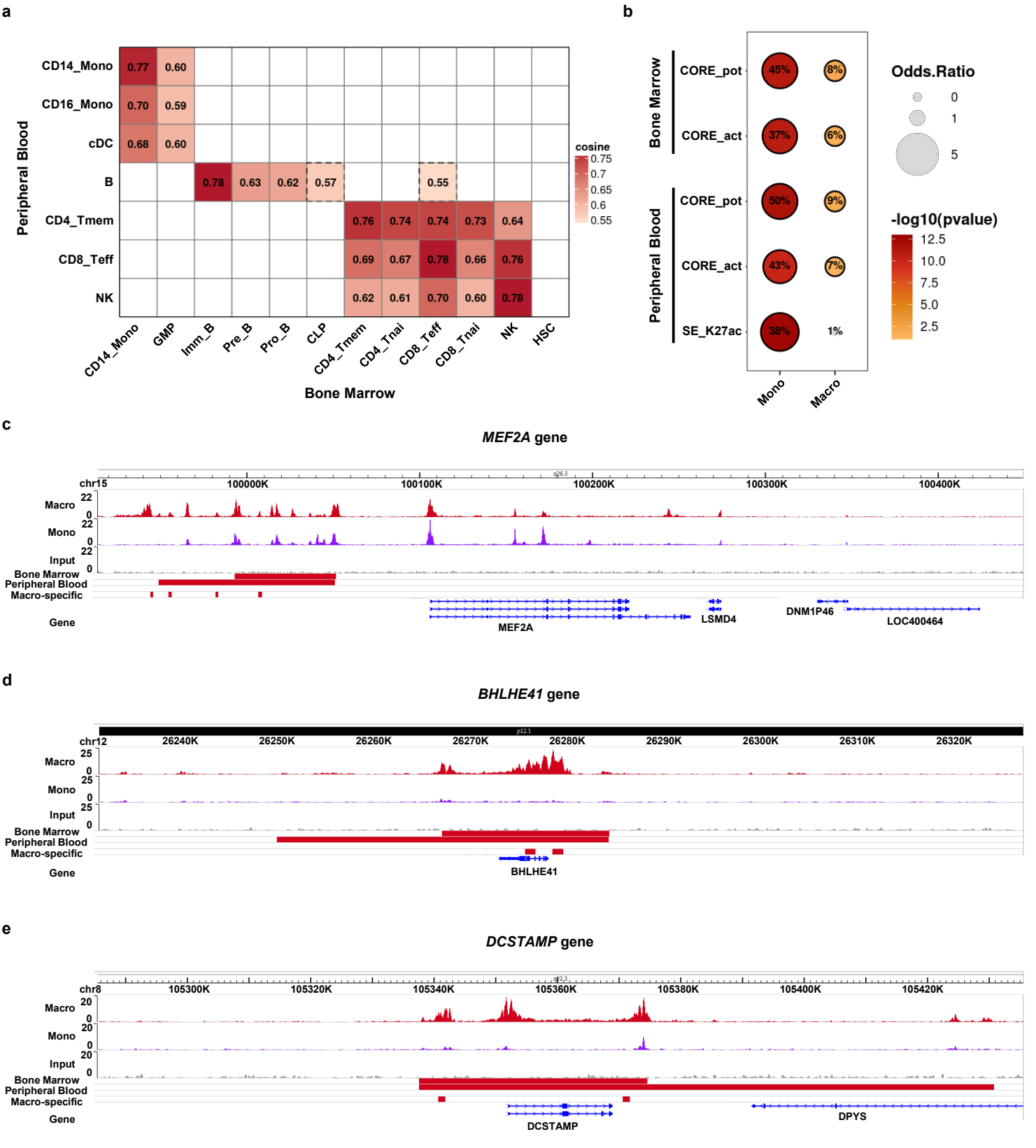
Extended Data Figure 5. Enrichment of fine-mapped CAD risk loci within CORE



Extended Data Figure 6. CORE constituents delineate lineage-specific epigenetic state

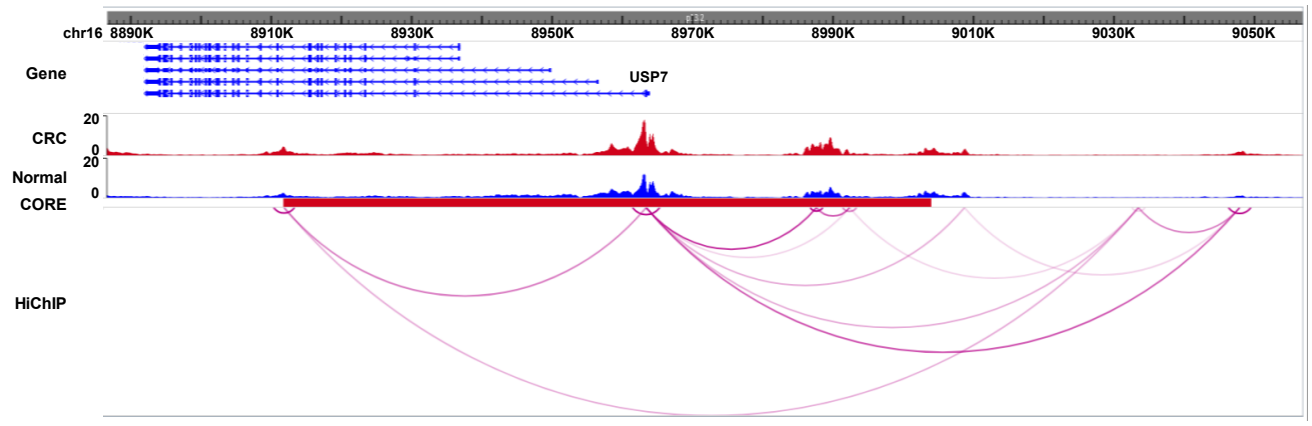


Extended Data Figure 7. enCORE captures hematopoietic lineage-matched BM-to-PB maturation trajectories

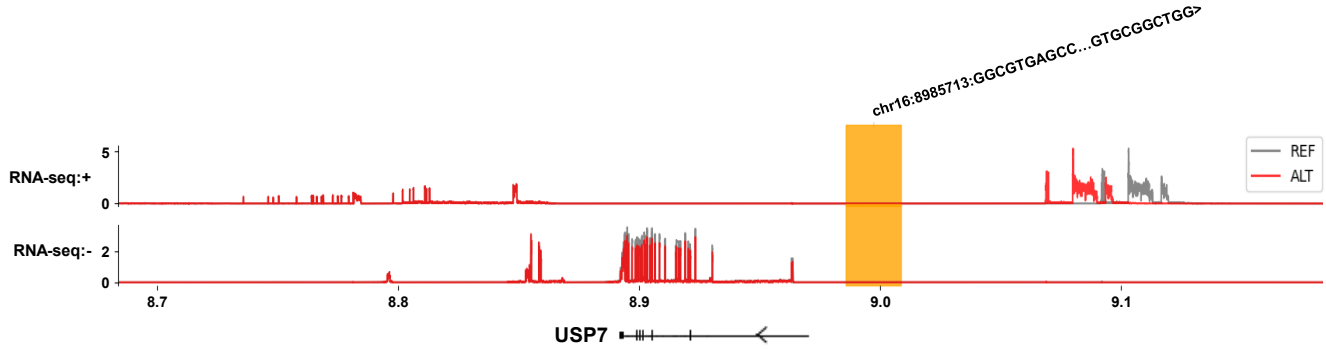


Extended Data Figure 8. ***In silico*** **knockout of a CORE-associated enhancer near** ***USP7***

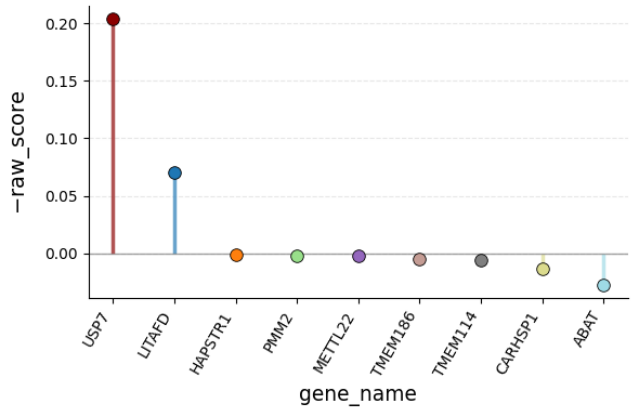
a



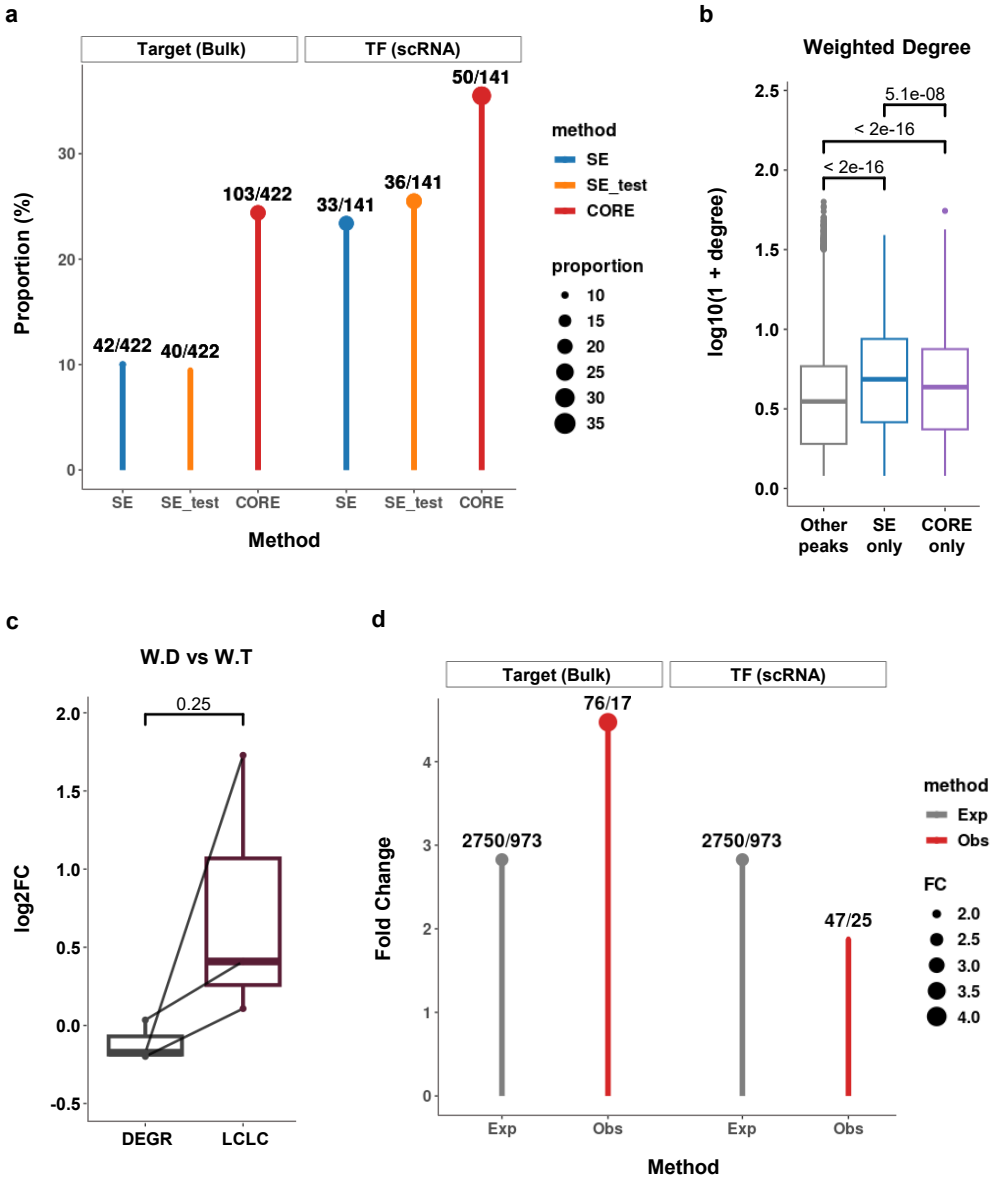
b



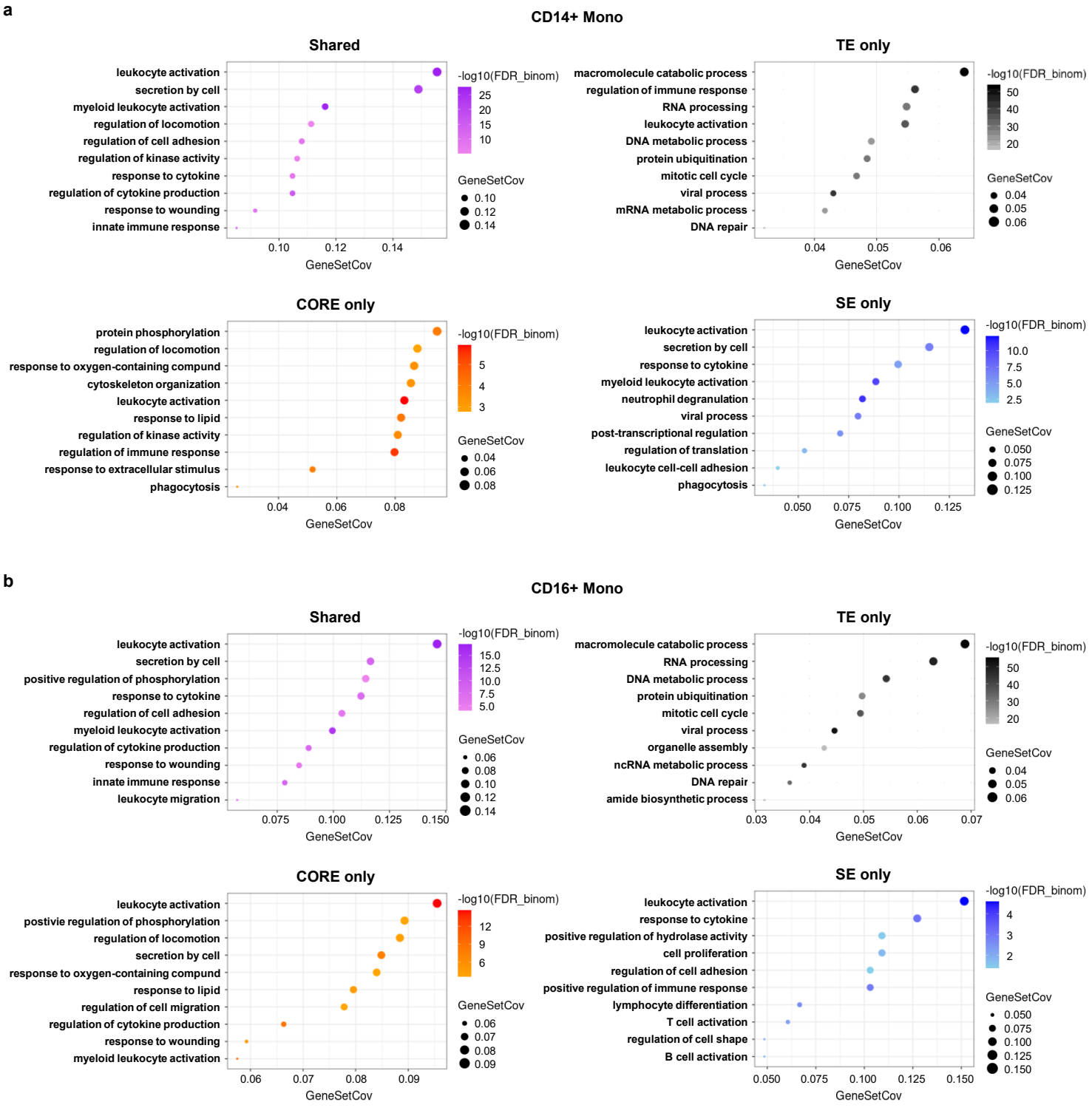
c



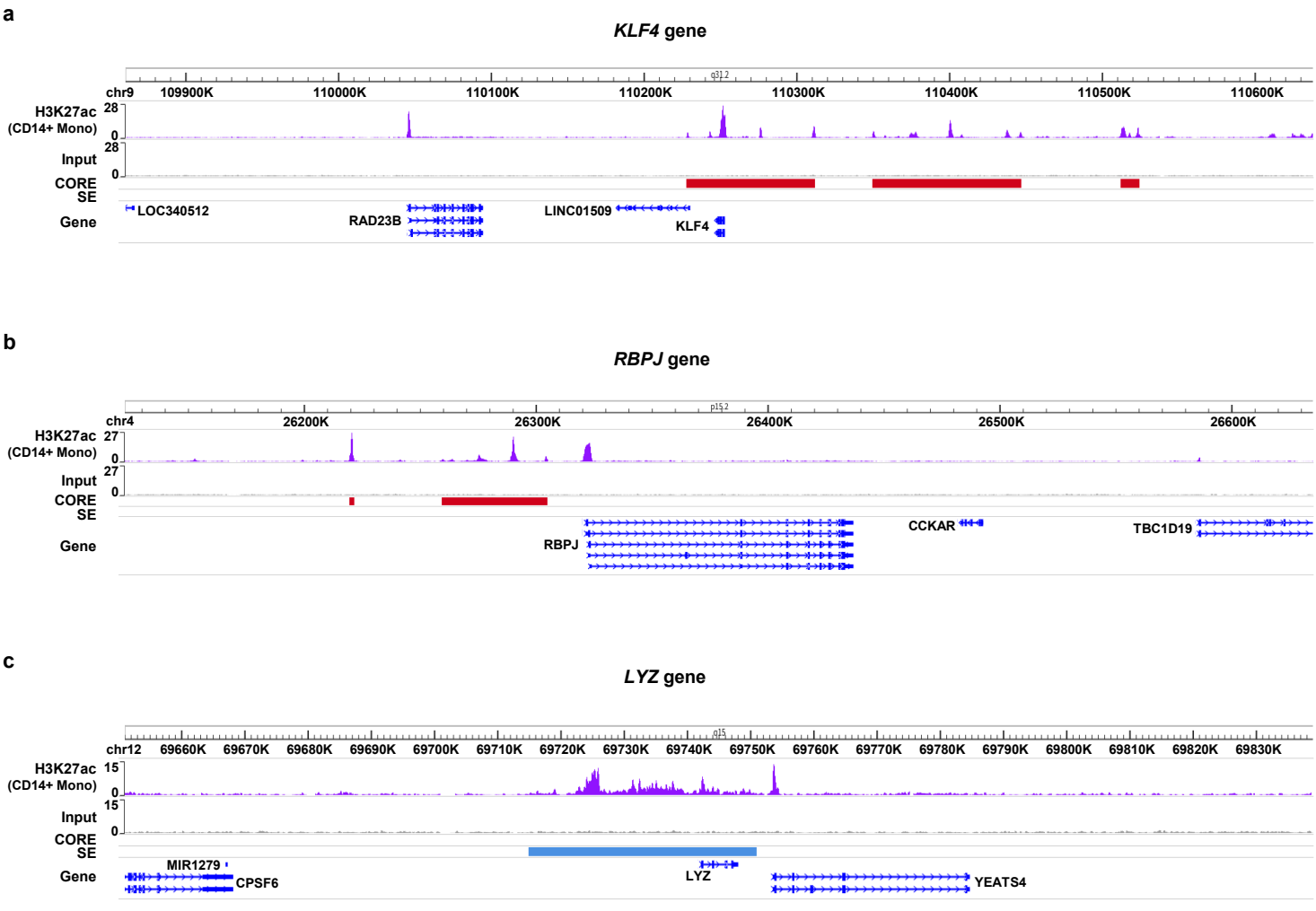
Extended Data Figure 9. Metastasis-related genes are enriched among CORE-associated genes and linked to distinct node-level network properties



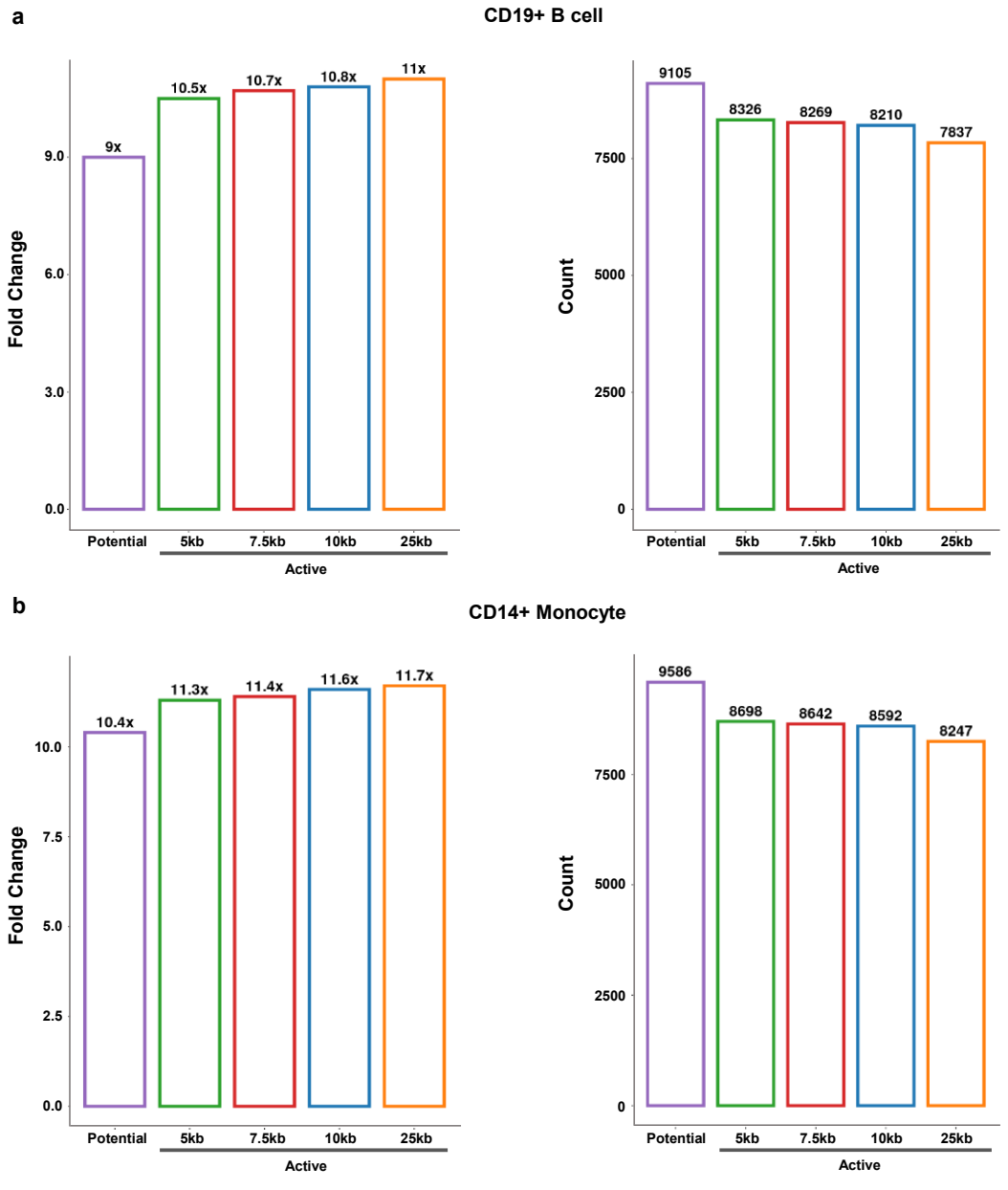
Supp. Figure 1. GO enrichment analysis of CORE in CD14+ and CD16+ monocytes



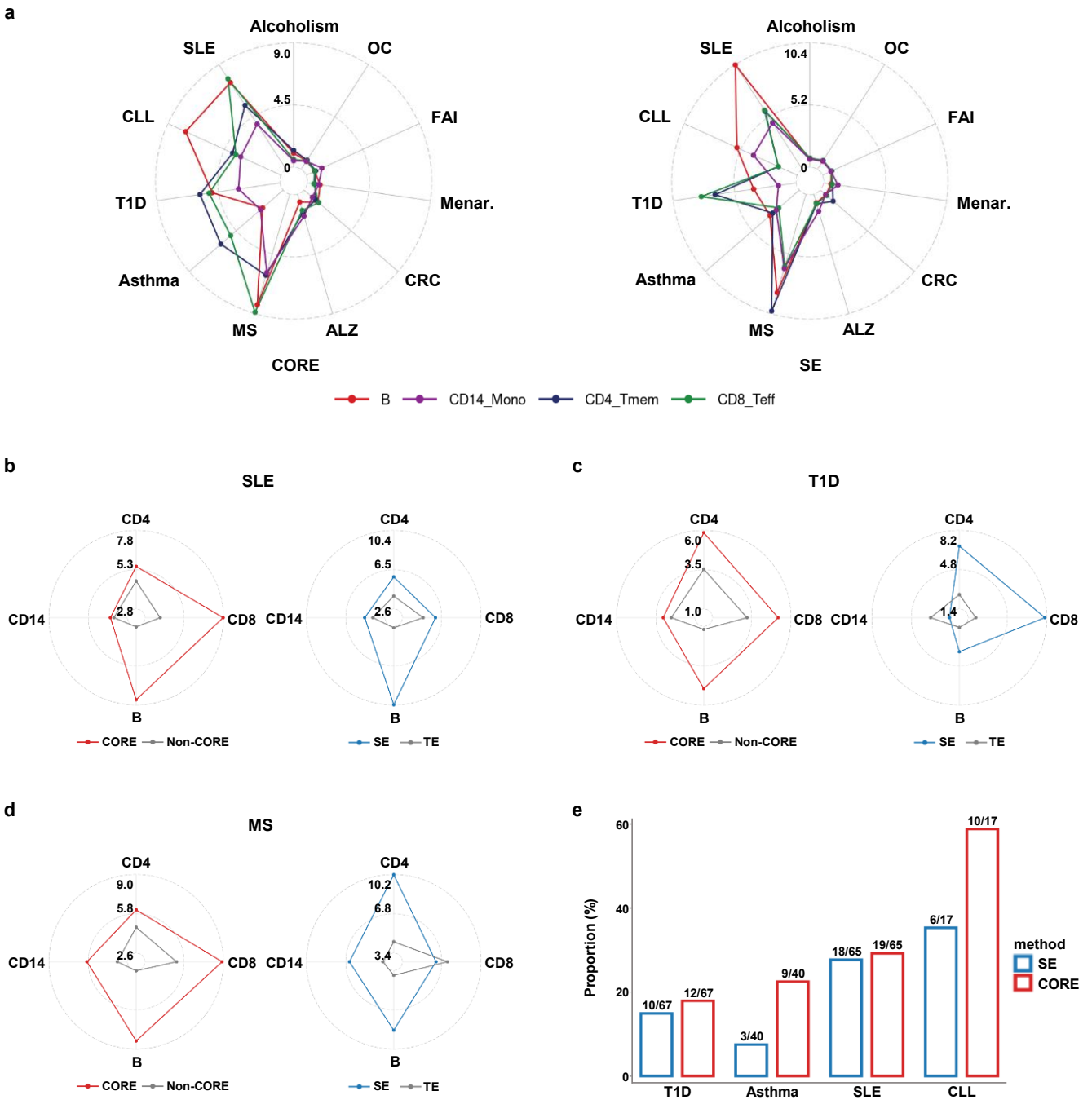
Supp. Figure 2. Track visualization of *KLF4*, *RBPJ*, and *LYZ* in CD14+ monocytes



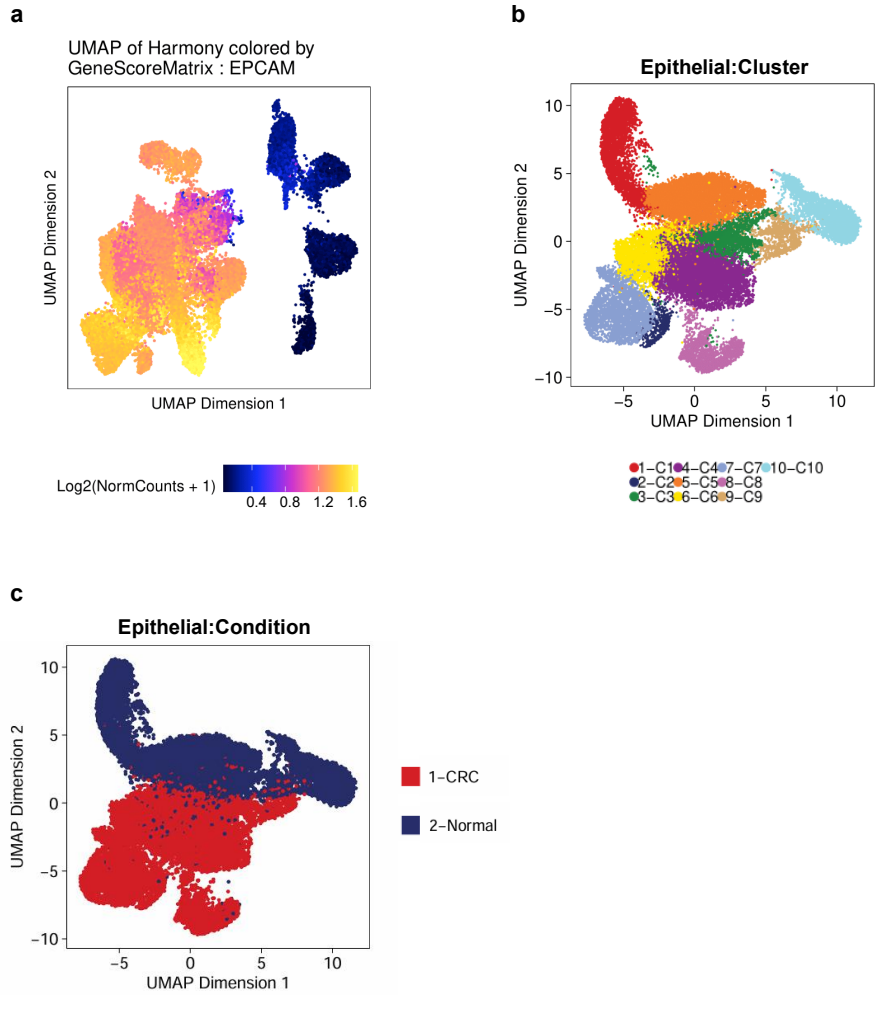
Supp. Figure 3. Active-option CORE shows a higher active-to-poised enhancer ratio than potential-option CORE



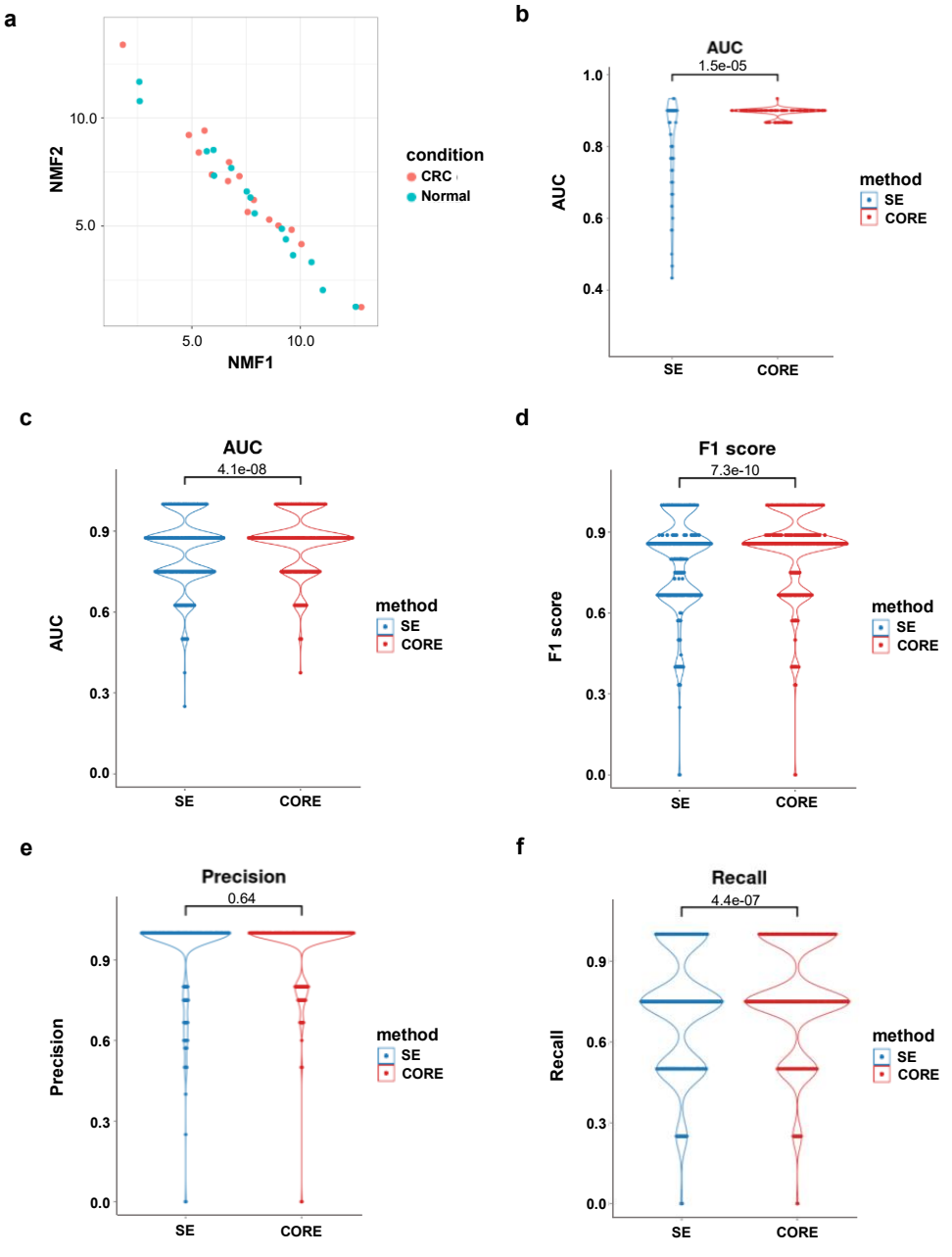
Supp. Figure 4. Immune-related disease GWAS variants are enriched in CORE (Hnisz *et al.*)



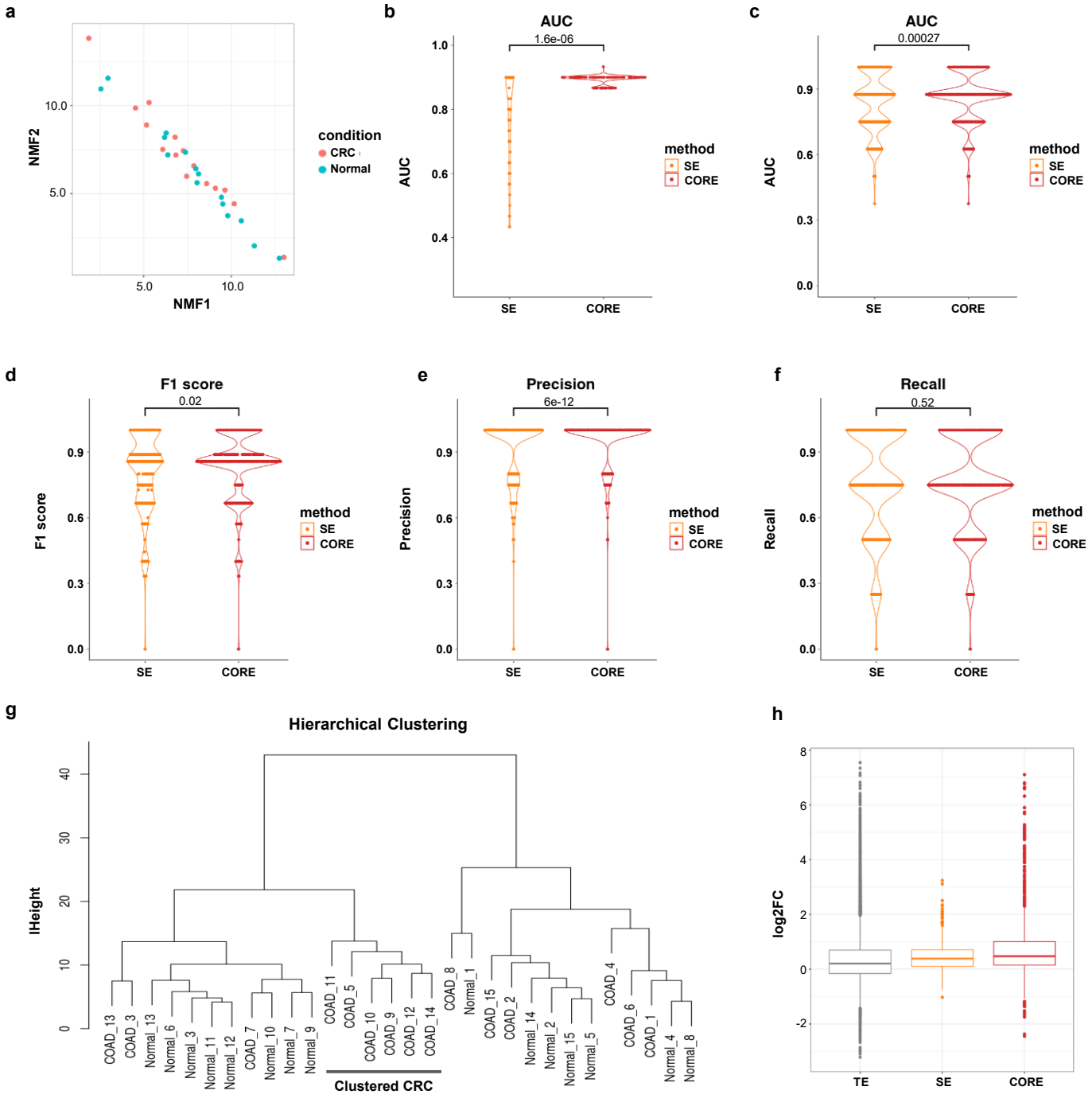
Supp. Figure 5. UMAP visualization and canonical marker expression in CRC epithelial cells



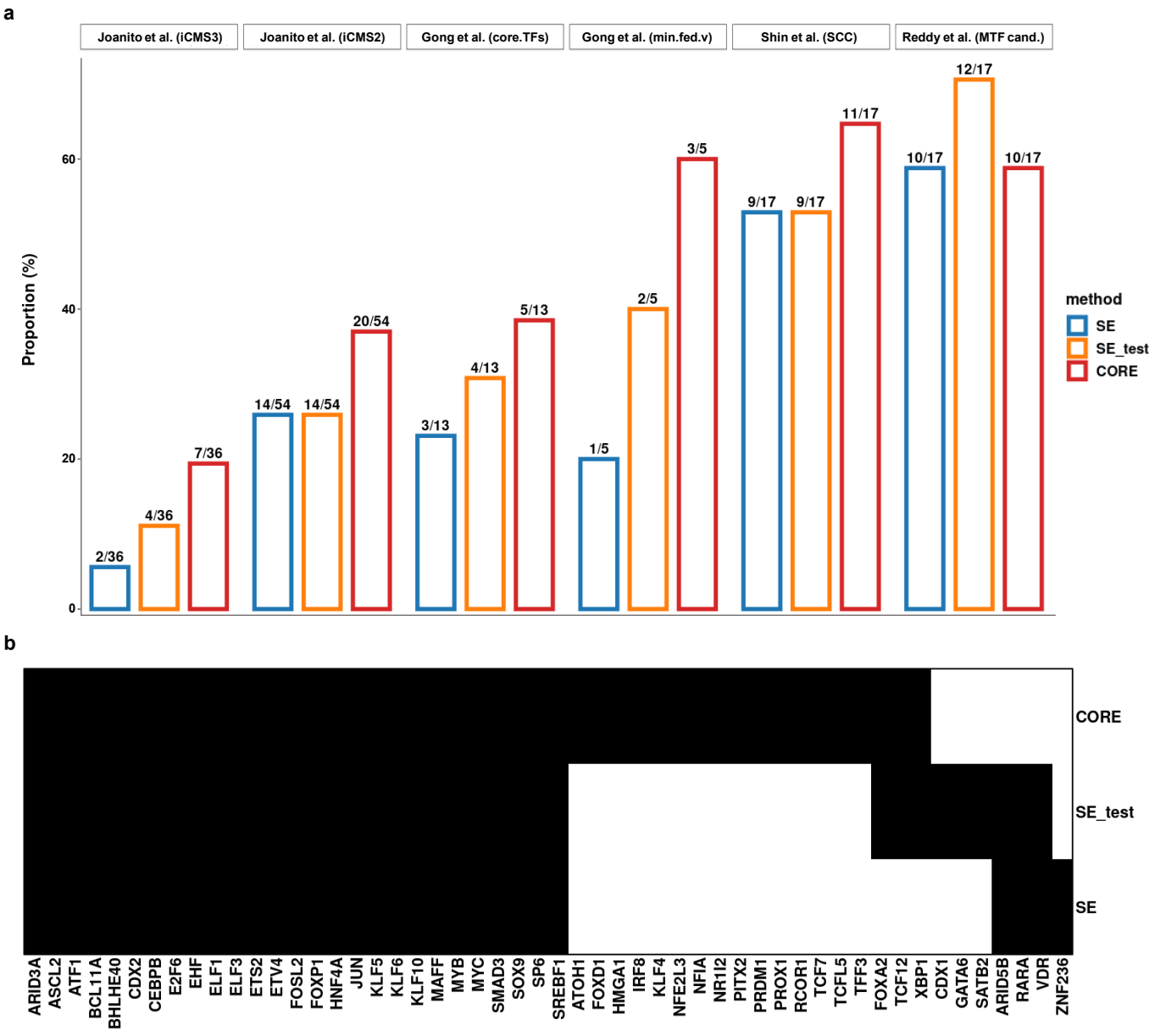
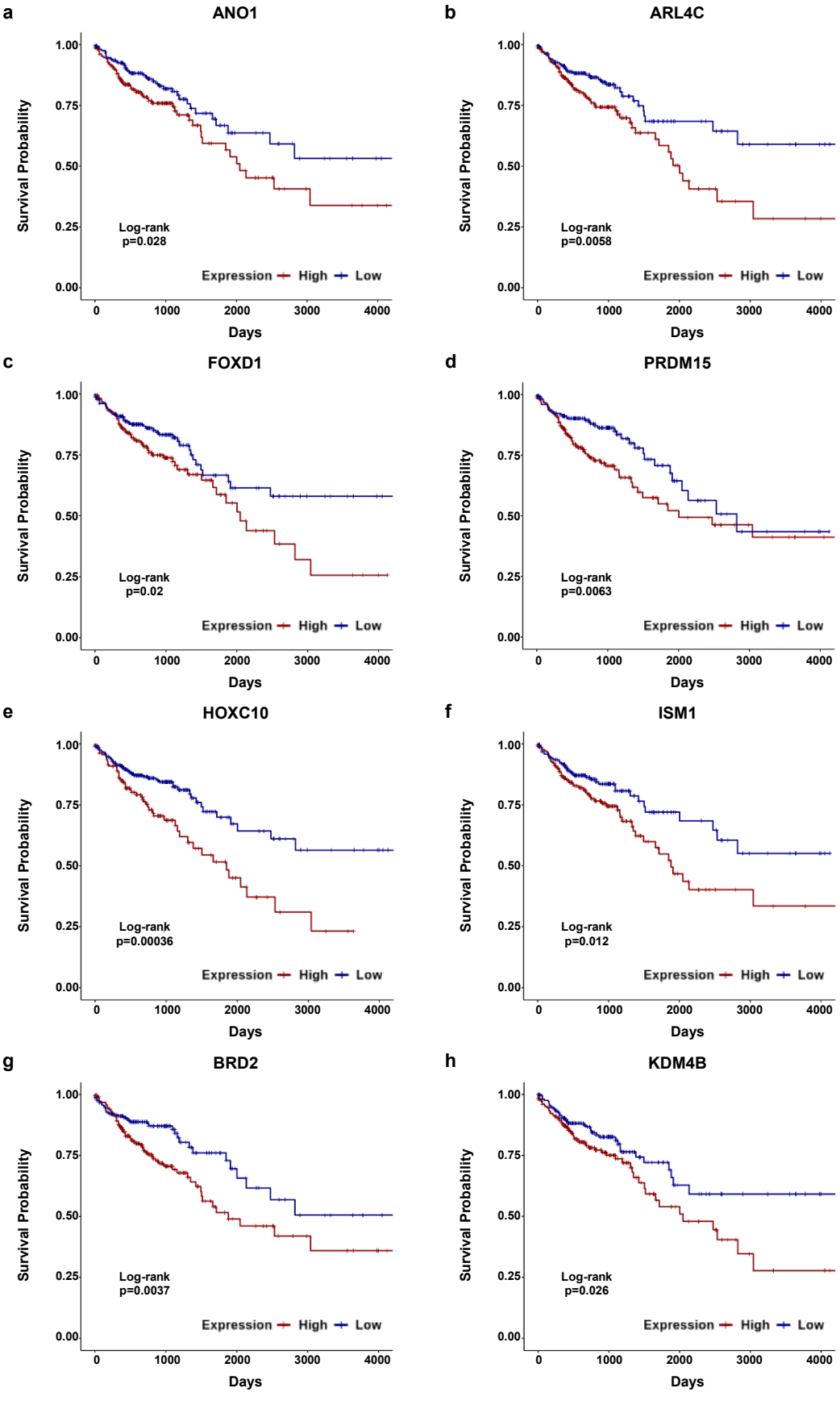
Supp. Figure 6. Comparison of classification performance between CORE and SE (GSE156613)



Supp. Figure 7. Comparison of classification performance between CORE and SE (GSE136889)



Supp. Figure 8. CRC master TF candidates are enriched among CORE-associated genes



Supp. Figure 9. Survival analysis for CORE-associated genes